Dispositius de control a la web

(Una topologia dels llocs webs a l'Estat Espanyol)

Òscar Coromina

Direcció: Emili Prado

Departament de Comunicació Audiovisual i Publicitat

Universitat Autònoma de Barcelona

Bellaterra, juliol de 2012

Nam et secundas res splendidiores facit amicitia

et adversas partiens communicansque leviores.

Ciceró

0. Índex

1. L'auca digital del Sr. Esteve (Introducció)

Navegar per la World Wide Web és avui una acció rutinària per milions de persones arreu del globus terraqui. En el transcurs d'aquesta activitat cada individu deixa traces digitals a mesura que es va relacionant amb cada un dels elements que integren la Web.

Imaginem-nos una escena carregada de quotidianitat:

El sr. Esteve arriba a la feina de bon matí i es posa a llegir les notícies del dia en una capçalera qualsevol de la premsa digital. Tot això ho fa mentre esmorza una magdalena i un cafè amb llet.

A l'ensems que entra en contacte amb l'actualitat de la crisi econòmica, el sr. Esteve va deixant la seva petja en aquells continguts als que dedica més atenció.

El nivell de granularitat de la informació recollida és remarcable: l'accés a una determinada pàgina web, el clic en algun dels enllaços presents a la pàgina, el visionat d'audiovisuals, la participació activa en forma de comentaris, la valoració d'una notícia, la compartició de contingut en xarxes socials d'Internet, la recepció d'impactes publicitaris, la resposta als mateixos impactes publicitaris i un llarg etcètera d'accions que van deixant rastres en ordinadors repartits per tots els racons del planeta.

Al sr. Esteve li recordaria els centenars de diminutes engrunes que van caient sobre la taula mentre mossega la magdalena i que delaten el seu esmorzar ric en sucre davant la pantalla del PC.

A més d'aquest registre d'activitat, l'ordinador que serveix aquesta informació té accés a certes dades que permeten identificar no tant al Sr. Esteve com a l'ordinador del sr. Esteve: l'adreça IP de la connexió, el sistema operatiu, el navegador utilitzat, la resolució de pantalla del dispositiu, etc.

Posem pel cas que el sr. Esteve, per accedir a un contingut determinat, ha hagut de passar per un registre d'usuari i, per tant, proporcionar algun tipus de

dada personal que l'identifica com a Ramon Esteve, home, 43 anys i resident a Torredembarra, província de Tarragona. Tota l'activitat del Sr. Esteve es pot associar a aquest perfil d'usuari

Com que aquesta escena passa en un moment àlgid del que s'ha batejat com a web 2.0, la interoperabilitat entre plataformes i serveis habilita el trànsit de dades simultani en diferents servidors de recollida de dades. Així, quan el Sr. Esteve decideix fer "un m'agrada" en una notícia de La Vanguardia que parla d'una iniciativa de l'alcalde de Barcelona per atraure turistes nòrdics a la ciutat, és presumible que el rastre del Sr. Esteve es registri tant en el servidor del diari digital com en els ordinadors de Facebook, relacionant-ho, en aquest cas, al perfil d'usuari del Sr. Esteve en aquesta xarxa social d'Internet[1].

La recreació d'aquesta escena ens permet il·lustrar un procés complex que transcorre de forma latent i no fàcilment perceptible mentre es navega per la World Wide Web. En aquest procés, les dades transiten d'ordinador client a servidor i acaben emmagatzemades físicament en terminals de diferents ubicacions geogràfiques.

És evident que si bé pel Sr. Esteve això no implica cap acte conscient, per aquelles entitats que recullen i emmagatzemen aquesta informació hi ha un interès en fer un ús determinat d'aquestes dades. En la majoria dels casos aquest ús estarà relacionat amb la consecució d'objectius com ara auditar el nombre d'usuaris del lloc web, adequar els impactes publicitaris a diferents públics objectius, vetllar perquè l'experiència d'ús sigui satisfactòria, millorar el rendiment dels processos de comerç electrònic, etc.

Aquest trànsit de dades no funciona únicament en sentit client-servidor[2]. Els servidors també poden enviar petits paquets de dades que es conserven en el

[1] Entenem que una xarxa social d'Internet (Social Network Site) consisteix un servei web que permet als individus: construir un perfil públic o semi-públic dins un sistema acotat, articular una llista d'altres usuaris amb els que es comparteix una connexió i veure i creuar la seva llista de connexions amb d'altres fetes per usuaris dins el sistema (D. Boyd i N. Elison, 2007)

[2] L'arquitectura client/servidor és un model de computació d'aplicació distribuïda en el que les tasques i càrregues de treball es divideixen entre els proveïdors de recursos i serveis –servidors- i els sol•licitants dels mateixos –clients- (wikipedia, 2012)

navegador web[3] durant un temps determinat. Aquesta capacitat és la que permet que determinats llocs web reconeguin un usuari recurrent i, fins i tot, conservin un registre de la seva activitat. Aquest mecanisme també és el que facilita que el Sr. Esteve no hagi d'introduir una i altra vegada el seu usuari i paraula de pas cada vegada que vol accedir a continguts i/o funcionalitats. Així, podríem dir que de la mateixa manera que els usuaris deixen rastres digitals, els servidors també poden deixar traces que evidencien la seva activitat.

Tot aquest moviment de bits amunt i avall d'Internet s'articula a partir del codi font dels documents que integren la World Wide Web. Dit d'una altra manera, les notícies (amb les seves fotos, els comentaris aïrats dels lectors i la publicitat de la casa d'apostes) són el resultat de d'un conjunt de sentències i instruccions escrites en un llenguatge de programació que s'executen en l'ordinador client.

A partir d'aquest relat es pot establir que les evidències de l'activitat digital del Sr. Esteve es troben tant en el seu ordinador personal (client) com en ordinadors d'arreu del món destinats a la recollida i anàlisi de dades. Al mateix temps, en el codi font de les pàgines que ha visitat el Sr. Esteve es fa evident no tant l'activitat del Sr. Esteve com el fet que aquesta activitat ha estat registrada amb una finalitat determinada.

Aquest treball de recerca es proposa analitzar el codi font de llocs web per trobar evidències d'aquest procés latent de recollida de dades i, en la mesura del possible (i també del factible), identificar actors, establir usos potencials i cartografiar els espais de la web tenint en compte el grau de monitorització al que està sotmès l'usuari quan els visita.

[3] Ens referim al software que permet accedir i visualitzar els continguts de la world wide web.

2. Marc Teòric

Segons el centre de terminologia de la llengua catalana (TERMCAT) un codi font és un conjunt de sentències escrites en llenguatge de programació, que un sistema informàtic processa per tal de poder-les executar (Colomer et al., 2003). Aquesta expressió és assimilable a la del castellà "código fuente", l'anglesa "source code" o les franceses "code d'origine" i "code source". En tots aquests casos, trobem la paraula codi acompanyada d'una altra que l'hi atribueix ser l'origen, la font o, dit d'una altra manera, el desencadenant d'un procés que es produeix en un ordinador.

Si intentem analitzar amb més detall les característiques d'aquest objecte digital, trobem que el codi font és la forma textual del codi de programació que editen els programadors informàtics. Conforma la primera part del procés de desenvolupament en el que s'indiquen les instruccions i els processos que ha de seguir un ordinador per arribar a un objectiu computacional determinat (Berry, 2011).

Aquesta forma textual es compon de caràcters alfanumèrics i símbols gràfics que conformen un arxiu que és un codi no compilat i no executable d'un programa informàtic. A diferència del llenguatge màquina que es basa en el codi binari, a la capa del codi font és fàcil trobar expressions que equivalen a paraules de la llengua anglesa. Resumint, diríem que és un conjunt d'instruccions escrites en una capa superficial dels llenguatges de programació.

El codi font és, per tant, una forma d'expressió abstracta que pren el llenguatge humà com a referència per descriure una operació concreta que, a través de la compilació posterior, es converteix en codi binari executable. És a dir, en una sèrie d'instruccions simples que es processen en un ordinador (Krysa i Sedek a Fuller, 2008).

The Art of Computer Programming de Donald Knuth és un llibre de capçalera en l'àmbit de les ciències de la computació. En aquest manual es proposa un símil entre la programació i les receptes de cuina que ens resulta d'utilitat per a

situar conceptualment el codi font. Els algoritmes dels llenguatges de programació, igual que les receptes, proveeixen un mètode: un conjunt de procediments formals prèviament definits que han de portar-se a terme en ordre per desenvolupar una tasca en un nombre finit de passos (Knuth, 2006).

Seguint amb el símil de la recepta de cuina, podem dir que el codi font és una descripció de les accions que es realitzen per assolir un objectiu determinat, els ingredients, les quantitats, el mètode de cocció, etc.

Estem davant, per tant, d'una forma textual composta d'expressions i formes sintàctiques que un individu competent en llenguatges de programació és capaç de comprendre i escriure d'una manera més fàcil que si es tractés de codi binari.

Aquesta descripció no només consta d'instruccions. També admet la inclusió de comentaris que no afecten la posterior execució de les instruccions. Normalment aquests comentaris serveixen per documentar la programació i incloure una descripció més detallada de les instruccions i funcionalitats. Fer-ne ús facilita la comprensió del codi per un altre programador en el supòsit que es requereixi una intervenció a posteriori o prendre aquestes línies de codi com a base per una nova aplicació.

Aquesta característica aporta una dimensió lingüística i fins i tot literària al codi font. A tall d'exemple reproduïm el poema "Listen" escrit en llenguatge de programació PERL per Sharon Hopkins, considerada un referent de la poesia PERL (Wall, 2000).

```
#!/usr/bin/perl

APPEAL:

listen (please, please);

open yourself, wide;
    join (you, me),
connect (us,together),
```

```
tell me.

do something if distressed;

    @dawn, dance;
    @evening, sing;
    read (books,$poems,stories) until peaceful;
    study if able;

    write me if-you-please;

sort your feelings, reset goals, seek (friends, family,
anyone);

        do*not*die (like this)
        if sin abounds;

keys (hidden), open (locks, doors), tell secrets;
do not, I-beg-you, close them, yet.

                            accept (yourself, changes),
                            bind (grief, despair);

require truth, goodness if-you-will, each moment;

select (always), length(of-days)

# listen (a perl poem)
# Sharon Hopkins
# rev. June 19, 1995
```

El poema de Sharon Hopkins ens serveix per il·lustrar una forma d'interpretar el codi font que clarament depassa l'estricta funcionalitat de donar instruccions a un ordinador.

A la vegada, ens permet enllaçar amb la Software Studies Initiative, una aproximació teòrica que, des de l'àmbit de les humanitats, proposa que el programari -el *software*- sigui considerat un objecte d'estudi i una disciplina pràctica per formes de pensar i treballar que històricament no "pertanyen" a l'àmbit de les ciències de computació. Aquestes noves àrees d'interpretació inclouen disciplines relacionades amb la cultura i els media des d'un punt de vista polític, social, filosòfic, estètic i artístic. D'aquesta manera, es proposa parar atenció en aspectes més o menys oblidats per l'aproximació dominant (Fuller, 2009).

Tot i que tradicionalment s'ha entès el programari com una entitat immaterial que es contraposa a la materialitat del maquinari -*hardware*-, des de la corrent dels Software Studies es parteix de la dimensió material del programari fixant-se en les característiques dels llenguatges i les interfícies, la sintaxi formulativa, els efectes que produeix i en com es relacionen aquestes pràctiques en l'articulació dels models socials actuals i en el desenvolupament de noves figures de coneixement. Fugint dels models de les ciències de computació i telecomunicació, es posa el focus en la virtualització, simulació, abstracció, feedback i processos autònoms (Fuller, 2009).

En una línia similar, *Philosophy of Software,* de David Berry, també aborda una reflexió sobre la importància i la transcendència dels estudis centrats en el programari ja que aquest s'ha convertit en una entitat vertebradora d'aspectes essencials de la nostra vida. L'autor para especial atenció en els fluxos[4] d'informació en temps real cap a ordinadors personals, servidors, telèfons, caixers automàtics... i en com aquests esdevenen un component clau en el desenvolupament de la quotidianitat.

Es diferencien els fluxos que es creen de manera conscient -el contingut generat per l'usuari- dels que es produeixen de manera inconscient a partir del

[4] Real-time streams

registre de l'activitat de l'individu. Segons apunta l'autor, per cada acció conscient hi pot haver més de 100 fragments d'informació que es registren i analitzen en processos invisibles a l'usuari.

Després, aquest flux porta les dades al "núvol" on s'agrega i s'analitza en temps real. A la vegada, dona peu a la creació de perfils d'usuari en base a variables de comportament que serveixen per habilitar nous fluxos des dels servidors per estimular respostes afectives i propiciar determinades accions (Berry, 2011).

El *software* articula bona part de les nostres accions diàries i la forma en que aquestes es desenvolupen es regula a través del codi.

Aquesta seria una de les premisses bàsiques que es desprenen d'un dels assajos de referència en l'estudi del codi realitzat des de l'àmbit de les ciències socials i més concretament de les ciències jurídiques: *Code and Other Laws of Cyberspace,* de Lawrence Lessig (1999).

La primera edició es va publicar al 1999 i és important situar aquesta obra en el seu moment històric. En aquells anys s'havia consolidat la idea d'Internet com un espai fora de l'abast dels poders regulatoris del món "real". Un escrit que ens situa en aquell context és *A Declaration of the Independence of Cyberspace*[5] publicat l'any 1996 per l'activista polític John Perry Barlow i que encara avui és un dels exponents més representatius d'aquesta visió utòpica i llibertària d'Internet:

> "Governments of the Industrial World, you weary giants of flesh and steel, I come from Cyberspace, the new home of Mind. On behalf of the future, I ask

[5] El terme ciberespai (cyberspace) té el seu origen en una novel·la de ciència ficció de William Gibson, autor de divisa del moviment ciberpunk, publicada l'any 1984. A Neuromàntic (Neuromancer) es descriu una realitat virtual paral·lela en un lloc denominat ciberespai. A tall d'anècdota convé citar que la novel·la és una de les fonts d'inspiració del film The Matrix. En l'anglès original del text de Gibson, The Matrix és una mena d'entitat del ciberespai a la que els navegants es connecten a través d'un implant cerebral.

> you of the past to leave us alone. You are not welcome among us. You have no sovereignty where we gather.
>
> We have no elected government, nor are we likely to have one, so I address you with no greater authority than that with which liberty itself always speaks. I declare the global social space we are building to be naturally independent of the tyrannies you seek to impose on us. You have no moral right to rule us nor do you possess any methods of enforcement we have true reason to fear.
>
> (...)"
>
> (Electronic Frontier Foundation, 2012)

És el propi Lessig qui, en el prefaci de la segona edició, presenta el seu treball com una refutació a la idea que l'entorn *online* pot mantenir-se al marge de la regulació del món *offline* (Lessig, 2009). Segons defensa aquest professor de dret constitucional, el codi és la pedra angular d'aquesta regulació ja que actua com una "constitució", entesa aquesta com una arquitectura que estructura i limita els poders socials i legals en base a la protecció d'un seguit de principis fonamentals.

La necessitat d'una major regulació a Internet s'ha de veure, no com un instrument per acotar els drets individuals, sinó com un mecanisme per protegir-los. La desregulació, en canvi, exposa a l'usuari al control d'altres individus, empreses i institucions (Lessig, 2009).

El mecanisme per articular aquesta regulació, defensa Lessig, és el codi, entès com el conjunt d'instruccions concretes que integren el codi font però també els protocols i normes en els que es basen les comunicacions a Internet. No es planteja com una opció, sinó com un fet que cal afrontar indefectiblement ja que és inherent a l'essència d'Internet i del software.

El codi i la possibilitat de canviar-lo, així com de modificar les estructures a un nivell superior, és l'instrument adequat per articular els drets dels usuari en

relació a la privacitat, accés a la informació i la llibertat d'expressió a Internet (Lessig, 2009).

L'absència d'arquitectures de control en els protocols que configuraven la Internet original, argumenta Lessig, no és un fet inamovible ja que el disseny de la xarxa –articulat a partir del codi i els protocols de comunicació- es pot canviar i, de fet, està canviant cap un entorn més regulat. Aquesta regulació, tanmateix, no es produeix a través del lideratge dels estaments legislatius tradicionals vinculats als estats i estructures supranacionals. Ans al contrari. Són les organitzacions comercials les que han propugnat noves arquitectures de control amb finalitats purament crematístiques.

El disseny original d'Internet presentava, segons aquest autor, tres "imperfeccions": l'accés sense credencials personals, l'omissió d'informació geogràfica i el flux de dades lliure -sense cap protocol que identifiqui els paquets d'informació i l'ús al que va destinat- (Lessig, 2009). L'evolució de l'arquitectura d'Internet ja ha establert noves capes de control. Des de força temps, l'assignació d'IPs es fa segons criteris geogràfics i els proveïdors d'accés a Internet estan obligats a conservar registres d'activitat dels usuaris durant un cert temps.

Un altre dels mecanismes que pal·lien (o aprofiten) aquestes mancances de la Internet original són les anomenades *cookies,* que apareixen l'any 1994 com una tecnologia d'identificació i monitorització amb l'objectiu d'incrementar els beneficis del comerç electrònic i millorar el servei als usuaris (Lessig, 2009).

El principi de funcionament de les *cookies* es basa en la gravació en l'ordinador client d'un petit arxiu que singularitza l'usuari assignant-li un codi d'identificació. D'aquesta manera es poden associar les diferents interaccions de l'usuari amb el contingut a un perfil determinat.

Aquest sistema dista encara de ser perfecte ja que en veritat només permet identificar l'ordinador i més concretament el navegador web (Firefox, Chrome, Internet Explorer, Safari, etc.). Però gràcies a les *cookies* es poden reconèixer usuaris recurrents, traçar el recorregut dels clics, associar-lo a patrons de comportament, etc.

Hi ha situacions que escapen a aquest mecanisme de control (varies persones poden tenir accés al mateix navegador, un individu pot utilitzar navegadors diferents, les *cookies* es poden esborrar i també bloquejar) però tot i així el volum d'informació recollida és remarcable i esdevé un actiu molt valuós pels llocs web.

Aquest petits arxius, que s'emmagatzemen en l'ordinador client, no tenen perquè contenir dades personals de l'usuari. Tanmateix, tècnicament és possible associar aquestes dades amb les que es generen per un altre via (per exemple, omplint un formulari de compra). En qualsevol cas, un dels aspectes més controvertits d'aquest sistema és la possibilitat tecnològica de que tercers es beneficiïn de la informació continguda a les *cookies*.

El pas definitiu de cara a la identificació de l'usuari i la pèrdua de l'anonimat és l'accés mitjançant perfils d'usuari que inclouen credencials personals. Al respecte, Lessig anticipa que l'establiment d'arquitectures de control a partir de credencials personals es produirà sobretot en xarxes tancades, propietat d'entitats privades.

Una altra aproximació especialment tinguda en compte en l'establiment del marc teòric ha estat la Digital Methods Initiative (DMI). L'epicentre d'aquest projecte acadèmic cal situar-lo a la Universiteit Van Amsterdam i al voltant de la figura de Richard Rogers. El seu principi fundacional es basa en el recurs a mètodes nadius digitals (*natively digitals*) en contraposició al que Rogers anomena mètodes digitalitzats (*digitized*).

S'entenen com a mètodes nadius digitals aquells que es basen en l'anàlisi d'objectes, continguts, dispositius i entorns "nascuts" a Internet. Com a mètodes digitalitzats es contemplaria la migració de les pràctiques estàndard de les ciències socials i les humanitats al nou mitjà (Rogers, 2009).

A partir d'aquesta distinció, podem establir que una enquesta *online*, l'observació dels hàbits de navegació o una entrevista usuaris d'Internet serien mètodes de recerca digitalitzats. Mentre que l'anàlisi dels resultats oferts per un cercador, les relacions entre els enllaços, les folksonomies, l'anàlisi del codi d'una pàgina web constituïren alguns exemples de mètodes nadius.

La proposta de Rogers i la DMI enllaça amb els plantejaments de Steve Jones que a l'any 1999 encetava un debat metodològic sobre la recerca a Internet amb el clar objectiu de qüestionar la utilització d'Internet com un objecte d'estudi independent del món real (Rogers, 2009). Igual que Lessig, Rogers deixa de banda la dialèctica evocadora del ciberespai per, en aquest cas, centrar-se en l'estudi dels "objectes" digitals: els *links*, els *websites*, els motors de cerca i les relacions que s'estableixen entre ells.

Per a portar a terme recerca empírica la DMI desenvolupa diferents aplicacions per recollir informació accessible a Internet, analitzar-la i visualitzar-la. El catàleg d'eines utilitzades és força extens i s'articula a partir d'un reguitzell d'objectes digitals prèviament definits. Al tractar-se de processos automatitzats, permet l'accés a una gran quantitat d'informació. D'aquesta manera s'obren les portes a treballar amb mostres força àmplies i, fins i tot, amb universos complets. Al mateix temps, això suposa una limitació en tant que "només" es pot obtenir dades d'informació accessible a la xarxa.

És important remarcar que la DMI no es presenta com un instrument per estudiar Internet, sinó que proposa el recurs als mètodes digitals per fer recerca empírica sobre diferents aspectes socials, culturals, polítics i econòmics del món que ens envolta.

En relació a l'activitat de recollida de dades de l'usuari que caracteritzàvem en l'apartat introductori, Richard Rogers utilitza l'expressió "*hit economy*" per descriure un model en el que anunciants, productes, serveis i proveïdors de continguts volen aparèixer i associar-se en els llocs més visitats (Rogers, 2002).

En posteriors assajos sorgits de la DMI s'identifica un nou model derivat de la popularització de les xarxes socials d'Internet: la "*like economy*". En aquest cas, es refereix a les aplicacions que permeten als usuaris mostrar la seva afecció -fer un m'agrada- a contingut que es troba fora d'aquestes plataformes socials. Aquesta activitat desencadena un seguit de fluxos de dades entre usuaris, propietaris de llocs web i les plataformes (Gerlitz. C i Helmond, A., 2012).

3. Objecte d'estudi

L'objecte d'estudi d'aquesta investigació són els processos de recollida de dades i monitorització l'activitat dels usuaris mitjançant el reconeixement de les expressions de codi que els identifiquen.

Per tal que aquest procés es realitzi de forma efectiva, és fa necessari l'ús d'una tecnologia que ho faciliti i a la que ens referirem com a dispositius de control.

La majoria d'aquestes aplicacions requereixen de la inserció d'uns fragments de codi en una pàgina web. Aquestes expressions desencadenen una acció que s'executa en el navegador de l'ordinador client i que habilita diferents fluxos de dades.

Tot i que tècnicament és possible desenvolupar una tecnologia pròpia per a realitzar aquesta tasca, l'existència d'empreses que comercialitzen software d'altes prestacions destinat a aquesta funció fa pensar que el recurs a aquesta opció és habitual entre els publicadors de la World Wide Web.

En termes de rendiment tecnològic aquesta opció té al seu favor que permet separar l'acció de servir la informació de la recollida de dades. És a dir, mentre el servidor dels continguts ofereix la informació, les dades de l'usuari s'emmagatzemen en un ordinador dedicat exclusivament a la monitorització. D'aquesta manera, s'alliberen recursos dels servidors dedicats a la publicació afavorint l'accés a la informació en millors condicions per part dels usuaris. Al mateix temps, els ordinadors i el software que recull les dades pot desenvolupar la seva activitat a ple rendiment.

El publicador, que prèviament ha contractat aquests serveis, pot accedir a informes que ofereixen un anàlisi detallat del comportament dels usuaris i en alguns casos automatitzar una resposta a certs indicadors.

En el present projecte de recerca l'anàlisi es focalitzarà en aquestes aplicacions desenvolupades per tercers i es realitzarà mitjançant la detecció d'expressions dins el codi font de la pàgina.

Les expressions reconegudes corresponen a un catàleg elaborat per Ghostery, una extensió que s'instal·la al navegador web i que fa visible en temps real la presència d'objectes dedicats a la recollida de dades del comportament dels usuaris i bloquejar-los. Ghostery és propietat de l'empresa Evidon, dedicada a les tecnologies de privacitat, que disposa d'un equip humà dedicat a l'elaboració un catàleg amb informació sobre els elements i les organitzacions que hi ha al darrera (Ghostery, 2012).

Instal·lant aquesta extensió s'obté accés a aquest catàleg des del mateix navegador web. Per cada un dels elements reconeguts trobem una descripció de funcionalitats i particularitats, afiliacions empresarials i informació de contacte. Al mateix temps, s'observa una classificació dels elements en 5 categories: *advertising* (elements per a la distribució publicitària), *analytics* (elements que faciliten l'anàlisi i la investigació per a publicadors web), *privacy* (avisos de privacitat i altres elements relacionats amb la privacitat), *trackers* (elements que no serveixen cap altre propòsit que la monitorització) i *widgets* (elements que articulen funcionalitat dins les pàgines com compartir o fer comentaris).

Aquesta categorització ens sembla poc consistent o, si més no, poc útil per la consecució dels objectius de la recerca. Per exemple, no acaba de quedar massa clar quin tret determina la diferència entre la categoria *trackers* i *analytics*.

Atenent-nos al que predica la Web Analytics Association podem definir l'Analítica Web com la medició, recol·lecció, anàlisi i elaboració d'informes a partir de dades d'Internet amb l'objectiu d'entendre i optimitzar l'ús d'un *website*. Fent un símil amb el mitjà televisiu podríem establir que l'analítica web és a Internet el que l'anàlisi d'audiències és a la televisió.

Hom pot establir els inicis de la indústria de l'analítica web l'any 1993 amb la creació de l'empresa Webtrends als Estats Units, tot i que no és fins el 1995 quan es comercialitza la primera aplicació d'analítica web: Webtrends. Webtrends era capaç d'analitzar l'activitat registrada en un servidor i mostrar-la en forma de taules i gràfics que permetien saber el nombre d'usuaris, dades sobre el seu perfil, continguts més populars, incidències tècniques i moltes

d’altres informacions sobre el comportament de l’audiència. (Digital Analytics Association, 2012)

Existeixen quatre tecnologies per recollir dades de la interacció dels usuaris amb els continguts online: *web logs, web beacons, Javascript tags i packet sniffing* (Kaushik, 2007).

Logs: Aquest sistema fou el primer mètode de recollida de dades d'Internet. Originalment va desenvolupar-se per registrar els errors que es produïen en els servidors web i ben aviat van veure ampliades les seves funcionalitats per proporcionar dades útils més enllà de l'aspecte purament tecnològic. El principi de funcionament és la generació d'un arxiu (log) on es registra cada una de les peticions realitzades en un servidor. La informació de les peticions crea una entrada en aquest arxiu on habitualment s'inclou la data, l'hora, l'adreça IP des d'on es genera la petició, el sistema operatiu i el navegador de l'usuari i l'objecte de la petició. Per analitzar la informació continguda en aquest registre es fa necessari l'ús d'aplicacions que tracten aquesta informació i la mostren de forma més intel·ligible.

Web beacons: L'aparició d'aquest sistema de medició està relacionada amb la utilització de *banners* com a reclam publicitari. El principi de funcionament és força simple i les seves funcionalitats limitades. Els *web beacons* són crides HTML a petits arxius d'imatge (normalment de 1x1 píxels) que es carreguen al mateix temps que el contingut de la pàgina web. Aquest petits arxius estan allotjats en altres servidors que es serveixen d’aquest element per mantenir un registre de la interacció.

Etiquetes de JavaScript: Els *tags* o etiquetes de JavaScript són a data d'avui el mètode més utilitzat per recollir l'activitat dels usuaris a la web. El sistema es basa en la inclusió de fragments de codi JavaScript que s'executen en el navegador web i que envia les dades de la sessió a un servidor que recull les dades. El sistema destaca per la seva precisió i té al seu favor que permet separar l'acció de publicar contingut de la recollida de dades. Dirigint aquests dos processos a ordinadors diferents s’afavoreix l’especialització i la optimització dels recursos informàtics.

Aquest factor ha estat decisiu en el desenvolupament de programari d'altes prestacions amb funcionalitats específiques que ha esdevingut germen de diferents iniciatives empresarials dedicades a l'emmagatzematge, recollida i anàlisi de dades i a l'elaboració d'informes.

Packet sniffing: Tècnicament es tracta del sistema més sofisticat per recollir dades i al mateix temps el menys popular. Consisteix en un element de hardware que s'interposa entre l'usuari i el servidor on es troba el *website* accedit. D'aquesta manera es registra la totalitat del flux d'informació entre els usuaris i el servidor.

Aquesta taxonomia de famílies tecnològiques ens sembla especialment útil per situar l'objecte d'estudi. Per una banda ens aporta la clau per entendre els criteris que expliquen la distinció que Ghostery fa entre *trackers* i *analytics*: Els *web beacons* corresponen a *trackers* i les etiquetes de JavaScript s'inclouen dins la categoria *analytics*.

Al mateix temps, ens fa prendre consciència de que l'anàlisi de codi font no ens servirà per detectar activitats de monitorització basades en l'anàlisi dels *logs* i *packet sniffing*.

Un repàs a la llista d'elements que integren la categoria "ad" de Ghostery revela diferents aplicacions relacionades amb els *adservers*, una família d'eines utilitzades per a la distribució d'impactes publicitaris a la web. Les dades no només flueixen en sentit servidor-client. L'activitat de l'usuari es recull per tal d'identificar-ne el perfil sociodemogràfic i les preferències electives. Aquesta informació és tinguda en compte en el mecanisme de distribució dels impactes a fi i efecte d'optimitzar l'impacte i el rendiment econòmic. (Sankuratripati, 2006).

La categoria *privacy* engloba un seguit d'elements als que s'atribueix un vincle amb la gestió i el control de la privacitat. En l'exploració superficial dels elements adscrits a aquest apartat afloren aplicacions per recollir i emmagatzemar informació del perfil de l'usuari alhora que eines que llancen avisos a l'usuari cada vegada es produeix un flux d'informació susceptible de comprometre la privacitat.

Finalment, la categoria *widget* inclou un seguit d'eines de natura molt diversa que tenen per finalitat articular fluxos de contingut en diferents plataformes. Dins d'aquest conjunt trobaríem diferents aplicacions relacionades amb el fenomen de la web 2.0 o web social. Utilitzem aquestes expressions per referir-nos a l'evolució de la web cap a un entorn més participatiu i col·laboratiu pel que fa a la generació de contingut (Beer, 2009). La majoria consistirien en botons per mostrar afecció a un contingut i compartir la informació publicant-la en blogs i xarxes socials d'Internet però dins aquesta classificació també hi trobem sistemes de sindicació de continguts i funcionalitats afegides al navegador.

Val la pena remarcar que tècnicament les xarxes socials no depenen de la participació activa de l'usuari per habilitar el flux de dades. El mecanisme de funcionament del *Like Button* de Facebook pot ser utilitzat per llegir la *cookie* que es genera quan un usuari accedeix a aquesta Xarxa Social i registrar la navegació i el comportament de l'usuari fora de la plataforma (Roosendal, 2010).

4. Objectius

L'objectiu de la investigació és topografiar l'ús d'aplicacions desenvolupades per tercers amb la finalitat de monitoritzar i recollir de dades dels usuaris a la web.

Per assolir aquest propòsit es fa necessari analitzar el codi font d'una mostra de llocs web amb la finalitat de detectar elements que fan palesa aquesta activitat.

En funció dels elements trobats podrem inferir quin ús fan els publicadors de continguts de les dades dels usuaris i amb quin propòsit.

Un objectiu secundari és el desenvolupament d'una metodologia que permeti l'anàlisi de la presència dels dispositius de control a la web. Per aquest motiu s'utilitzen diferents eines que automatitzen la detecció dels elements, la recollida de dades i l'anàlisi de mostres relativament àmplies.

5. Preguntes de la investigació

El present projecte es proposa respondre a les següents preguntes:

- Quins són els dispositius de controls més utilitzats?

- Amb quin propòsit es recullen les dades?

- Existeix alguna relació entre les eines utilitzades i les característiques dels llocs web?

6. Mostra

La mostra s'elabora a partir de la creació de llistats que agrupen diferents llocs webs a partir de la seva afinitat temàtica i/o de continguts. En alguns casos la mostra serà equivalent a l'univers, en d'altres seran la representació de diferents espais a la World Wide Web a partir d'una caracterització prèvia.

Els universos escollits són: els portals dels ajuntaments de les capitals de província de l'Estat Espanyol, els de les universitats que integren la CRUE (Conferencia de Rectores de Universidades Españolas), capçaleres de la premsa digital espanyola, plataformes de comerç electrònic de bitllets d'avió i continguts pornogràfics a la xarxa. Amb aquesta selecció es cerquen espais de contrast a partir de la confrontació entre model públic i privat, llocs webs informatius i transaccionals, models de negoci, oferta de productes i serveis, etc.

Al mateix temps, es para atenció en sectors empresarials i institucionals en els que la presència a la xarxa es considera estratègica per a dur a terme les seves activitats i que, per tant, haurien d'haver assolit un cert grau de maduresa tècnica i professional.

Per elaborar els diferents conjunts es recorre a diverses fonts i tècniques:

En el cas dels ajuntaments s'elabora una cerca manual dels portals dels ajuntaments de les capitals de província espanyoles fins a obtenir les 50 urls.

El llistat corresponent a la CRUE s'extreu del *website* de l'associació on es relacionen 75 llocs web.

La confecció dels mitjans de premsa digital és més complexa ja que es decideix prioritzar els principals actors de la premsa digital en termes d'audiència i difusió. Per aquest motiu es consulten els rànquings que ofereixen regularment OJD i EGM i, en detectar absències destacades en cada una de les fonts, se'n crea un de nou per procés de fusió. En el cas d'OJD es realitza una consulta a l'informe corresponent al mes de març de 2012 amb la finalitat d'obtenir el llistat de llocs web més visitats. D'aquest rànquing es seleccionen únicament les capçaleres informatives fins arribar a obtenir els llocs web de premsa digital

més visitats segons OJD. Paral·lelament, s'accedeix a l'informe de febrer-març de 2012 de l'EGM i se n'extreu el llistat dels 25 llocs amb més visitants únics. Seguint el mateix criteri que l'aplicat en el cas anterior s'exclouen aquells llocs web que no poden ser considerats capçaleres informatives. S'obté una mostra integrada per 66 urls.

Al no tenir accés a cap informació oficial que permeti identificar els llocs webs de referència en termes d'activitat econòmica, es decideix utilitzar una tècnica que afavoreix els actors més competents a l'hora d'obtenir trànsit d'usuaris procedent dels cercadors. Els llocs webs de la mostra s'obtenen a partir dels resultats oferts pel cercador Google en relació a una cadena de cerca determinada. Per evitar la contaminació dels resultats a partir de preferències personals de l'investigador s'utilitza un navegador sense cap *cookie* emmagatzemada i sense cap sessió d'usuari oberta.

S'introdueix la cadena de cerca "*vuelos baratos*" a la url google.es desmarcant qualsevol filtre de cerca, indicant la opció idiomàtica "*español*" i obviant els resultats de Google Instant. Es seleccionen les urls corresponents als 100 primers resultats de cerca orgànica, s'eliminen duplicats i es sotmeten a un procés de transformació en dominis (www.domini.com) per obtenir les adreces de les pàgines principals. La operació resulta en una mostra de 96 urls.

El mateix procediment es repeteix amb la cadena de cerca "*sexo gratis*" obtenint una selecció de 98 urls.

Com a resultat d'aquest procés obtenim una mostra de 385 urls categoritzades en cinc espais diferents que es pot consultar a la taula 1 de l'annex.

A les taules 1, 2, 3, 4 i 5 que s'inclouen en l'annex de taules i figures es detallen les urls que integren els diferents espais.

Taula 6: Distribució de la Mostra per espais

Media	Capitals	Crue	Sexe	Vols	Total
66	**50**	**75**	**98**	**96**	**385**

7. Mètodes i tècniques

Per a la recollida de dades que acreditin l'existència de dispositius de control als llocs web seleccionats es recorre a l'ús d'una eina desenvolupada per la DMI (Digital Methods Initiative) de la Universiteit Van Amsterdam.

L'eina, batejada amb el nom de Tracker Tracker, detecta expressions que identifiquen un ampli ventall de d'eines utilitzades per la monitorització dels usuaris. El repertori inclou software d'analítica web, botons d'interacció social, servidors de publicitat i altres elements de tercers presents al codi HTML de llocs i pàgines web. Segons es descriu al directori d'eines de la DMI, el Tracker Tracker es capaç de reconèixer la presència més de 900 dispositius de diferents.

Aquest catàleg de rastrejadors procedeix d'una altra aplicació anomenada Ghostery que s'instal·la com una extensió del navegador web que va informant a l'usuari de la presència d'aquests dispositius alhora que es va navegant.

El principal avantatge de Tracker Tracker front a Ghostery és que permet la introducció d'una llista de fins a 100 urls diferents i, de forma automatitzada, extreu diferents arxius amb les dades tabulades per al seu posterior anàlisi.

En l'apartat de desavantatges, cal explicar que Tracker Tracker es basa en una llista d'elements que es correspon amb el catàleg de dispositius reconeguts per Ghostery al febrer de 2012. Ghostery actualitza permanentment el seu catàleg de manera que molt probablement hi haurà hagut incorporacions posteriors a aquesta data que no podran ser detectades.

Sospesant aquestes limitacions s'opta per l'ús Tracker Tracker ja que sembla més adequat per l'anàlisi de les mostres seleccionades.

El procés d'extracció de dades es realitza sense incidències i en un únic dia. L'eina utilitzada permet extreure per cada un dels espais tres arxius: El primer és un arxiu csv on es relacionen les urls amb els elements detectats en forma de matriu. El segon, també un arxiu csv, ofereix, a més de les urls amb els elements associats, informació addicional sobre les característiques dels

elements detectats que inclou una categorització a partir de la seva funcionalitat. El tercer és un arxiu .gefx, un format específic per al programa de visualització de dades Gephi on les urls i els rastrejadors s'assimilen a nodes i les connexions entre ells a arestes.

Una primera exploració dels resultats permet observar que les urls on no s'ha trobat cap rastrejador no consten als arxius resultants. Es procedeix a la detecció manual d'aquesta circumstància i es crea una nova taula que reflecteix els llocs web on el resultat de la detecció ha estat nul.

En un intent de donar més consistència a la categorització realitzada per Ghostery, s'assimilen les categories *analytics* i *trackers*, en una nova: analítica web.

8. Limitacions

Com hem explicat anteriorment, el procés de recollida de dades es realitza mitjançant una aplicació que automatitza el procés de detecció. El recurs a aquesta eina permet ampliar considerablement les dimensions de la mostra i identificar un ampli ventall d'elements.

Tot i així, cal tenir en compte que l'eina utilitzada es basa en el reconeixement d'una sèrie d'expressions d'un catàleg prèviament definit que recull les aplicacions desenvolupades per tercers més utilitzades.

Això suposa dues limitacions destacables: no es detectaran aquelles aplicacions de tercers que no figurin en el catàleg i no es detectaran els sistemes d'elaboració pròpia per monitoritzar els usuaris.

Creiem que aquest fet pot afectar especialment als actors locals i és la detecció d'una absència destacada la que porta a aquesta reflexió. En el procés d'elaboració de la mostra de l'espai de mitjans detallàvem la inclusió de les capçaleres informatives auditades per OJD. Es dona el cas que en el codi font figuren expressions que delaten la presència del seu dispositiu de control però no apareixen referides en els arxius resultants del procés de recollida de dades.

Un aspecte, observat a posteriori del procés de recollida de dades i anàlisi, és l'aparició d'elements duplicats en una mateix lloc web. Examinant amb deteniment els arxius proporcionats per Tracker Tracker es conclou que això es deu a l'existència de dues expressions diferents en la forma però relacionades amb una mateixa eina. Poden haver-hi diferents motius que expliquin aquest fet: la coexistència de versions diferents del software de monitorització, una mala praxi en la inserció de les expressions dins el codi font, l'existència de diferents mòduls d'un mateix software, el recurs a diferents perfils de medició, etc.

Valorant l’impacte d’aquest fet en els resultats i veient que no afecta a les preguntes i hipòtesis plantejades, s’opta per mantenir la presència recurrent del mateix dispositiu de monitorització en un mateix lloc web.

També en l’apartat de limitacions, cal tenir en compte que només s’analitzen les pàgines d’inici dels llocs web seleccionats. Això significa que l’anàlisi es centrarà en un nivell superficial de la navegació i que existeix la possibilitat que en un segon nivell es pugui estar utilitzant altres eines de monitorització que habiliten nous fluxos de dades. Seria el cas d'un mitjà que decideix ubicar els *widgets* socials a les pàgines de les noticies i deixar la portada neta d'elements externs.

Per bé que les expressions que delaten els dispositius de monitorització són una part integrant del codi HTML existeix la possibilitat de que aquestes expressions no es trobin al arxiu HTML de les pàgines web analitzades sinó en un arxiu independent. Tal pràctica es deu a un mètode de programació que consisteix en agrupar totes les instruccions de JavaScript en un altre arxiu minimitzant així les interferències entre diferents llenguatges de programació. És possible que l'ús d'aquesta tècnica de programació impedeixi la detecció dels elements per part del Tracker Tracker. Una altra mètode que impediria la detecció d'elements seria l'encriptació del codi font però és una pràctica molt i molt residual.

La presumpció que el software desenvolupat per tercers és el recurs habitual per recollir i analitzar dades sobre l'usuari es basa en l'experiència professional de l'autor d'aquest projecte de recerca en l'àmbit de la comunicació i el màrqueting digital. Tanmateix, és necessari remarcar que les troballes realitzades no pretenen descartar en cap cas el recurs a tècniques i tecnologies que escapen de l'escrutini automatitzat del codi font com seria el cas de la medició per *logs* i *packet sniffers*.

Els resultats obtinguts han de ser interpretats com el que són: una fotografia amb un enquadrament molt ampli però que no representa la globalitat. Recordant la paradoxa d'Aquiles i la tortuga, diríem que la globalitat ens sembla un objectiu inabastable i que ens conformem en *tendir* a la representació global.

9. Antecedents

Poc a poc, l'analítica web va esdevenint un tema recurrent en la literatura científica. Una interrogació sobre investigacions recents a les bases de dades accessibles aporta diferents aproximacions sobre la matèria.

Un dels treballs que val la pena destacar el treball de Katzuo Nakatano i Tao-Tao Chuang: *A Web Analytics Tool Selection Method: an Analytical Hierarchy Process Aproach* (2011), que proposa una metodologia per a l'elecció d'eines d'analítica web que s'adeqüin a les diferents necessitats de les organitzacions.

Una altra publicació que indaga sobre els factors que marquen la continuïtat d'ús dels serveis d'analítica web és la publicació *Determinants of continuous usage intention in web analytics services* (2009) de Jaesung Park, JaeJon Kim i Joon Koh. Mercès a una enquesta realitzada a 152 organitzacions s'estableix que tant la dependència com la satisfacció juguen un paper clau per explicar la fidelitat cap a una eina o altra.

En una línia similar però en aquest cas més centrada en l'ús de l'analítica web al servei dels objectius de negoci trobem *A practical Evaluation of Web Analytics* de A. Phippen, L. Sheppard i S. Furnell (2004). A través d'aquest estudi (realitzat en una empresa multinacional de línies aèries) es mostra el potencial de l'analítica web quan es posa al servei de l'estratègia empresarial a l'ensems que posa en relleu les dificultats en promoure en la organització la consciència del valor de l'analítica web per a realitzar l'activitat professional.

Més orientat a qüestionar la precisió del sistema de medició basat en l'ús de *cookies* apareix *How Google Analytics and Conventional Cookie Tracking Overestimate Unique Visitors* (2010) de Max I. Fomitchev. L'autor conclou que, de la mateixa manera que els sistemes de comptabilització d'usuaris a través d'IP, les *cookies* fallen en el propòsit de ser precises a l'hora d'identificar usuaris únics. Esgrimeix motius com el fet que una mateixa persona pot

connectar-se des d'ordinadors o navegadors diferents, l'acció dels antivirus, l'actualització de sistemes operatius i, fins i tot, el borrat manual per part dels usuaris. Tots aquests factors inflen el compte d'usuaris únics d'una manera substancial.

Més crítics, i preocupats pels aspectes relacionats amb la pèrdua de privacitat dels usuaris, K. Balachander i C.E. Wills., en el seu treball P*rivacy Diffusion on the web: A longitudinal Perspective* (2009), apunten a una tendència creixent en el volum de dades que es registren dels usuaris i com, al mateix temps, es redueix el nombre d'entitats que efectuen la recollida, registre i tractament de les dades. A més, examina les diferents tecnologies d'obtenció de dades i el nivell de detall i profunditat que es pot arribar a assolir.

La tecnologia utilitzada per a la distribució d'impactes publicitaris, els *ad servers*, també ha captat l'atenció de diferents estudis recents.

A The *Online Advertising Industry: Economics, Evolution, and Privacy* (2009) de David S. Evans es realitza una caracterització exhaustiva del negoci de la indústria publicitària a la xarxa i de la importància que ha adquirit en termes de penetració a la web i volum de negoci. A la vegada, desgrana els detalls de la relació simbiòtica entre contingut i publicitat *online* i els reptes que ha d'afrontar des d'un punt de vista d'estructura del sector, prestacions i com això pot derivar en una pèrdua de privacitat de l'usuari.

Orientat a analitzar i millorar el rendiment dels algoritmes de distribució, trobem *Efficient Online Ad Serving in a Display Adverstishing Exchange* (2011) de K. Lang et al. La recerca es proposa la millora en termes d'eficiència del software desenvolupat i utilitzat per l'empresa Yahoo a partir de tres eixos de funcionament: idoneïtat de la pàgina web i el publicador, idoneïtat de l'usuari que està visitant la pàgina web en qüestió i idoneïtat de l'anunci i l'anunciant.

En relació als mecanismes de compartició de continguts a les xarxes socials d'Internet i l'activitat de recollida de dades que se'n deriva també trobem alguns estudis recents.

A Facebook Tracks and Traces Everyone: Like This (2010) Arnold Roosendaal reflexiona sobre les implicacions que comporta la implementació del Facebook Like Button en els llocs web. Concretament, fa èmfasi en el fet que la inclusió

d’aquesta funcionalitat serveix per la generació de *cookies* als ordinadors dels usuaris i, a partir d’aquí, per a que Facebook disposi de més informació per a incorporar al perfil dels usuaris. Tot plegat, segons defensa l’autor, té importants implicacions en la privacitat de l’usuari en tant que el procés de recollida de dades es produeix encara que els usuaris no facin ús del botó en qüestió.

L’expansió dels “botons socials” de Facebook també és motiu de reflexió a The *Like economy – Social buttons and the data-intensive web* (2011) de Carolin Gerlitz i Anne Helmond. Aquesta publicació fa èmfasi en que, paral·lelament a la creació d’una web social, aquests botons esdevenen una activitat econòmica basada en la recollida intensiva de dades dels usuaris que denominen *Like economy*. Darrera de la promoció del contingut generat per l’usuari i la socialització a través de xarxa social, s’estaria creant una infraestructura on la participació dels usuaris serviria per alimentar la *Like economy*.

10. Resultats i anàlisi

Executat el procés d'extracció de dades en els 385 llocs web que composen la mostra, el recompte d'elements detectats aflora 1444 dispositius de control. No estem parlant d'actors diferents; molts dispositius es troben presents en diferents llocs web de forma simultània. Si tenim en compte aquest fet, identifiquem 78 dispositius diferents.

La presència d'elements en els diferents espais que integren la mostra es fa visible a la taula següent:

Taula 7: Nombre d'elements detectats

Media	Capitals	Crue	Sexe	Vols	Total
577	**82**	**149**	**297**	**339**	**1444**

Aquestes primeres dades ens permeten comprovar que la major concentració d'elements de dispositius de control es dona a l'espai dels mitjans digitals. A continuació, trobem els grups corresponents als espais de contractació de vols i continguts sexuals que presenten xifres properes. Finalment, seria en els portals dels ajuntaments i de les universitats de la CRUE on s'habilitaria un menor quantitat de fluxos de dades dels usuaris. Tot i així, al tractar-se de mostres de mides diferents s'ha d'incorporar aquesta variable per fer l'anàlisi.

Dividint els elements detectats pel nombre de llocs web que integren les diferents mostres obtenim la ràtio d'elements per lloc web que es pot veure a la Taula 8.

Taula 8: Ratio d'elements per lloc web

MEDIA	CAPITALS	CRUE	SEXE	VOLS	TOTAL
8,74	**1,64**	**1,98**	**3,03**	**3,53**	**4,01**

S'aprecia clarament que a l'espai dels mitjans l'activitat de recollida de dades es produeix amb més intensitat. Més del doble que a l'espai de sexe i vols i 4 vegades més que als ajuntaments i les universitats.

Les taules anteriors no reflectirien un altre indicador que també ens dona pistes sobre la intensitat de control: el nombre de llocs web on el procés de detecció ha estat nul. S'identifiquen un total de 65 llocs web que segueixen aquesta pauta. Repartits per espais, els resultats nuls es distribueixen de la següent manera:

Taula 9: Llocs webs amb resultats nuls

Media	Capitals	Crue	Sexe	Vols	Total
0	**16**	**19**	**16**	**14**	**65**

Igual que abans, es fa necessari ponderar aquestes xifres tenint en compte la mida de la mostra.

Taula 10: Percentatge de resultats nuls

Media	Capitals	Crue	Sexo	Vuelos	Total
0%	**31,25%**	**25,33%**	**16'32%**	**14'58%**	**16,8%**

A la taula 10 es reflecteix, en forma percentatge, els llocs web sense dispositius reconeguts. Seria agosarat afirmar que es tracta de llocs on l'usuari pot

navegar lliure de control ja que també pot ser degut a l'ús de tecnologies indetectables amb el mètode utilitzat per a la recollida de dades. El que sí podem afirmar amb menor marge d'error és que sembla molt i molt complicat eludir la monitorització quan es llegeix la premsa digital. L'usuari hauria de tenir la precaució de bloquejar l'emmagatzematge *cookies*. I aquesta mesura de protecció activa de la privacitat tampoc seria suficient per escapar de dispositius de control basats en *web beacons, packet sniffers* o *logs* que, recordem, no podem detectar amb Tracker Tracker.

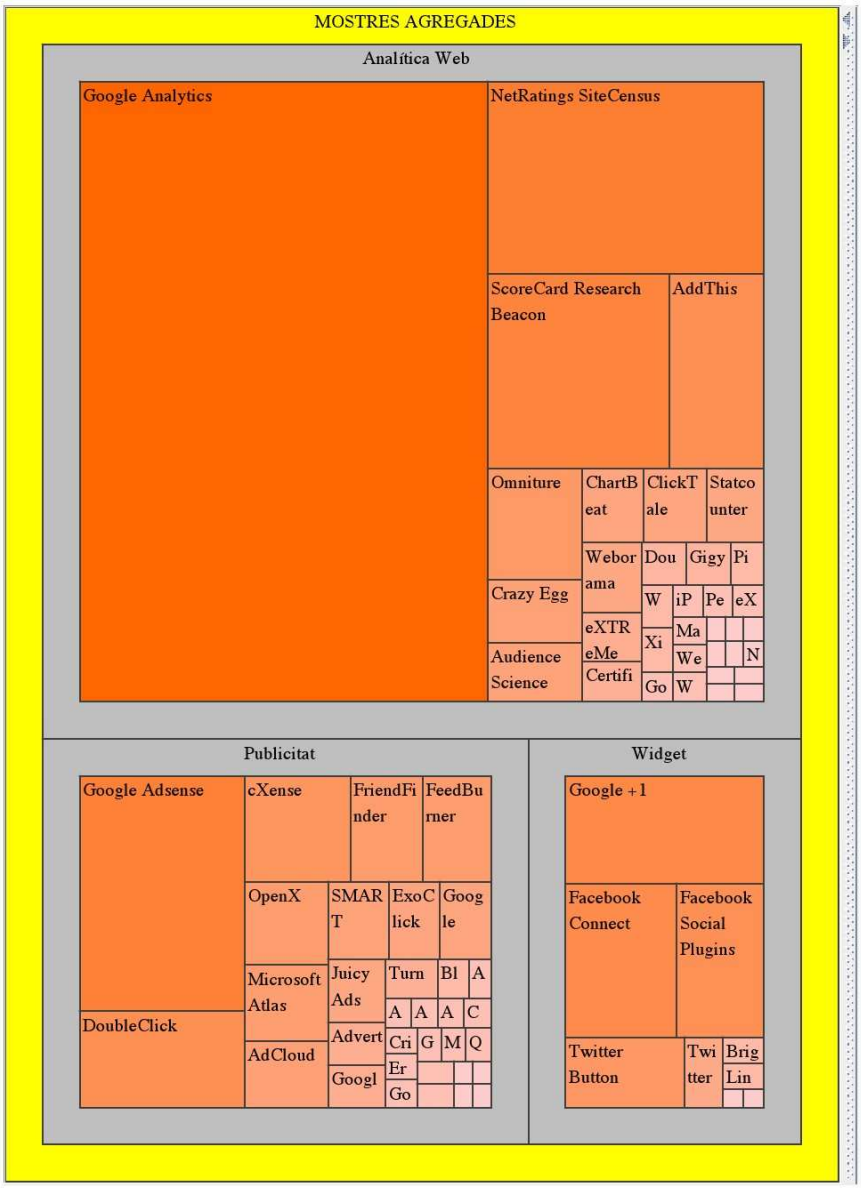
La llista de dispositius detectats i la seva recurrència es pot consultar a les taules 11, 12, 13, 14, 15 i 16 que s'inclouen a l'annex de taules i figures

Anàlisi topològica amb *treemaps*

Per a l'anàlisi topològic dels diferents espais que integren la mostra farem servir un gràfic en forma de *treemap* (un tipus de visualització permet la representació de dades a partir d'estructures jeràrquiques). En el cas que ens ocupa només és recorre a dos nivells: la categoria (Analítica Web, Publicitat, *Widget* i Privacitat) i el nom dels elements.

En les figures que venen a continuació (1, 2, 3, 4, 5 i 6) els diferents elements detectats es mostren com a quadrilàters que al mateix temps es troben dins dels quadrilàters de les categories. La superfície que ocupen els polígons es proporcional al nombre de vegades que l'element apareix a l'espai analitzat. El color també és més intens en funció de la recurrència.

Figura 1: Treemap dels espais agregats



La figura 1, que representa totes les mostres agregades, revela que l'activitat de recollida de dades està molt dominada per l'analítica web (i en aquest apartat Google Analytics és un actor destacadíssim) seguida de la publicitat i els *widgets*.

Un altre fet que es desprèn d'aquesta visió panoràmica és l'absència d'aplicacions de les que Ghostery inclou en la categoria privacitat. Sembla doncs, que aquesta tipologia de dispositius no s'utilitza en cap dels llocs de la mostra.

Observem també que la pràctica totalitat d'elements adscrits a la categoria *widget* serveixen per habilitar fluxos de dades amb xarxes socials d'Internet a través de la compartició de continguts i altres funcionalitats. Els actors més importants en aquest apartat són, per ordre: Google +1, Facebook Connect, Facebook Social Plugins i Twitter Button. Encara que Google +1 és l'aplicació amb una presència més recurrent, no s'ha de perdre de vista que Facebook articula la seva posició a partir de dos elements diferents. Resulta si més no curiós comprovar que encara que Google plus gaudeixi d'una base d'usuaris sensiblement inferior respecte Facebook i Twitter, ha obtingut un èxit gens menyspreable a l'hora de convèncer als publicadors web perquè incloguin aquest *widget* als seus llocs.

Els dos elements vinculats a la plataforma Facebook realitzen funcions diferents. Facebook Connect identifica l'usuari relacionant-lo amb les dades del seu perfil a la xarxa social i permet l'accés dels propietaris del website visitat a certa informació del perfil (facebook developers 2012). Facebook Social Plugins, en canvi, activa diferents funcionalitats orientades a compartir, recomanar i comentar el contingut a la xarxa social.

Fora de la categoria de *widget*, trobem un element que per les seves característiques creiem que caldria incloure en aquesta classificació però que apareix ubicada dins el rectangle de l'analítica web. Es tracta d'AddThis, una aplicació que facilita la compartició de continguts a les xarxes socials d'Internet i que, a jutjar per la mida del seu rectangle, és d'un ús bastant extens. Podem atribuir aquesta ubicació a un cert grau de discrecionalitat de la categorització de Ghostery utilitzada per Tracker Tracker. Val a dir que en el catàleg

actualitzat de Ghostery aquesta mala categorització ha estat corregida i ja consta com a *widget*.

Abans de passar a l'anàlisi particular de cada un dels espais, volem fer èmfasi en el fet que les plataformes propietat de Google Inc. (Google Analytics, Google Adsense, Double Click i Google plus) són especialment rellevants en termes de penetració i posició dominant. De fet en totes les categories (publicitat, analítica web i *widget*) encapçalarien el rànquing de recurrència.

Si ens fixem en l'estructura dels *treemaps* segmentats detectem patrons similars en la configuració d'alguns dels espais.

Figura 2: Treemap espai CRUE

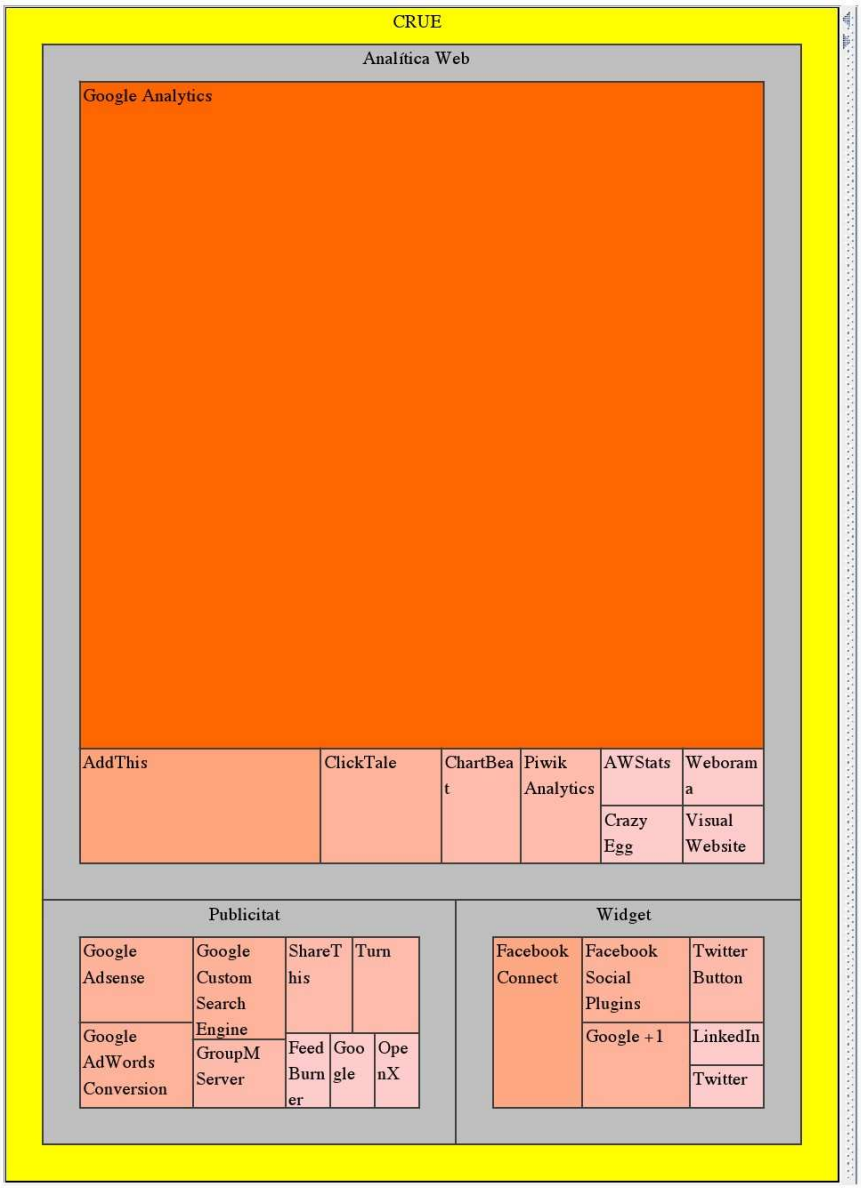
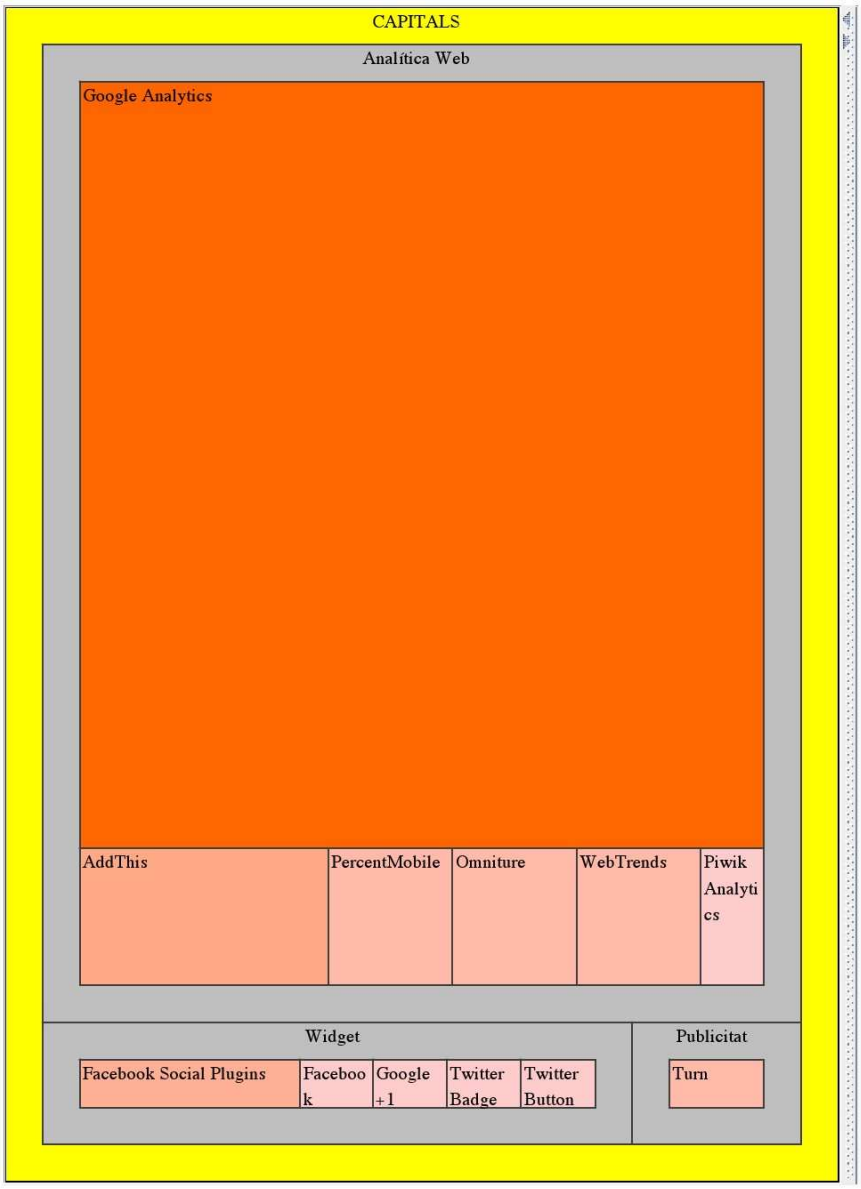


Figura 3: Treemap espai capitals de provincia



És el cas dels llocs web de la CRUE (fig. 2) i de les capitals provincials (fig. 3), on el rectangle que inclou les eines d'analítica web ocupa una gran superfície. Dins d'aquest polígon, es fa evident l'hegemonia de Google Analytics com a sistema de medició.

En aquests dos espais i de forma molt més testimonial, trobem també la presència d'elements adscrits a les categories de *widgets* i de publicitat (essent l'activitat publicitària molt més escassa als llocs web de les capitals provincials). Si tinguéssim en compte Addthis com a *widget* es faria més visible la preponderància dels *widgets* en relació a la publicitat.

Tot i que, com hem dit, el pes de la categoria publicitat és molt reduït, no deixa de cridar-nos l'atenció que aquests elements apareguin en llocs webs que, a priori, no es serveixen del model publicitari per a generar ingressos. Consultant la informació que ofereix Ghostery, altra volta detectem inconsistències en el criteri d'elaboració del catàleg que Tracker Tracker utilitza per extreure les dades i que han estat rectificades en la versió actual del catàleg. És el cas de Google Custom Search Engine (que facilita la inclusió de la funcionalitat del cercador Google a qualsevol lloc web) que apareix llistat dins la categoria publicitat i en la versió actual ja s'inclou dins l'apartat de *widgets*. Un altre cas que val la pena comentar és el de Google Adwords Conversion, un element que serveix per auditar amb més precisió els resultats obtinguts a les campanyes publicitàries realitzades al cercador Google. Si bé és cert que està relacionat amb l'activitat publicitària, ens sembla necessari deixar palès que la

presència d'aquest element no significa que es dugui a terme cap activitat d'explotació publicitària dels llocs web.

Figura 4: Treemap espai vols

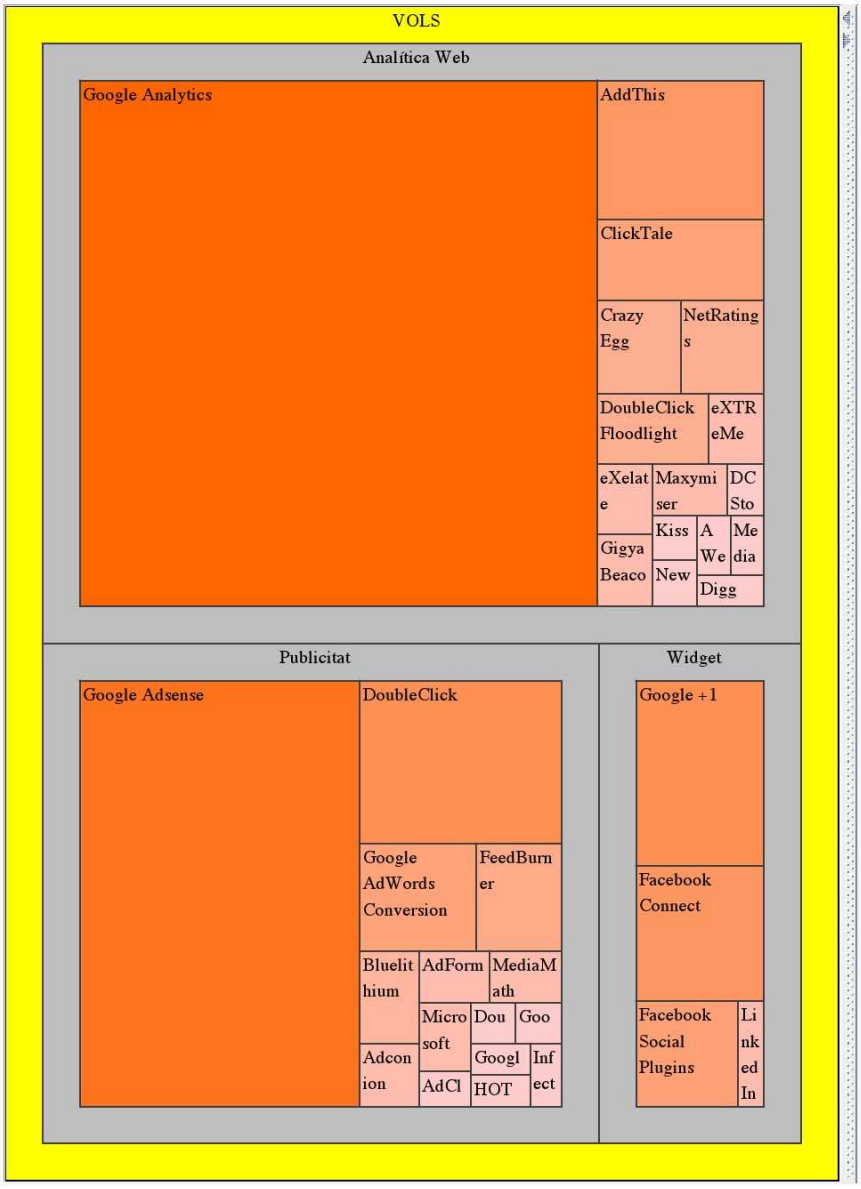
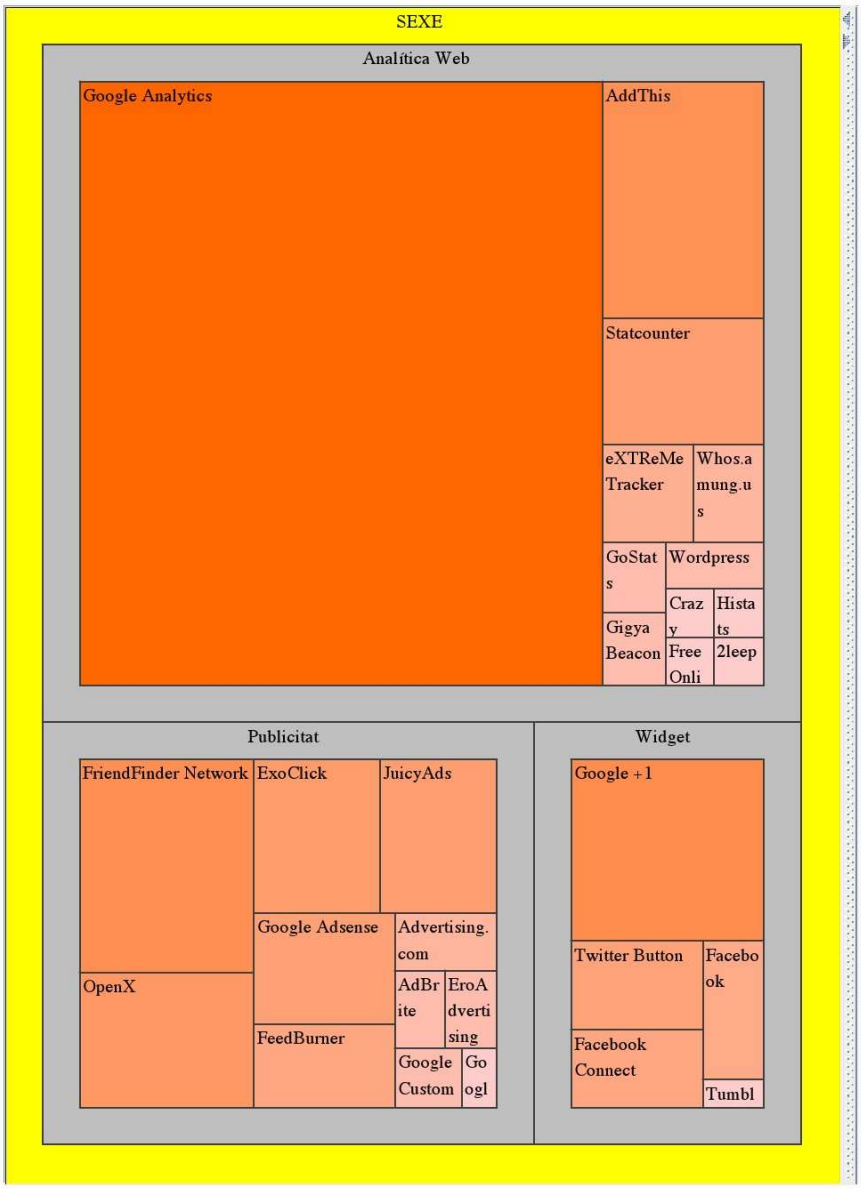


Figura 5: Treemap espai sexe



En els llocs relacionats amb la contractació de vols (fig. 4) i entreteniment per adults (fig. 5) també identifiquem un patró similar. Tot i que les eines d'analítica web també són majoritàries, es percep un presència més significativa dels elements que integren les categories de publicitat i *widgets*. En tots dos casos els elements inclosos en la categoria publicitat són clarament més nombrosos

que els *widgets*. En aquest ocasió sí que ho podem relacionar amb l'explotació publicitària dels llocs web.

A més d'aquesta distribució més compensada de la superfície detectem una major fragmentació, cosa que evidencia l'ús d'un repertori més variat d'eines. Tot i així, Google Analytics segueix essent, de lluny, el dispositiu més utilitzat i sembla bastant clar que és la opció estàndard de medició.

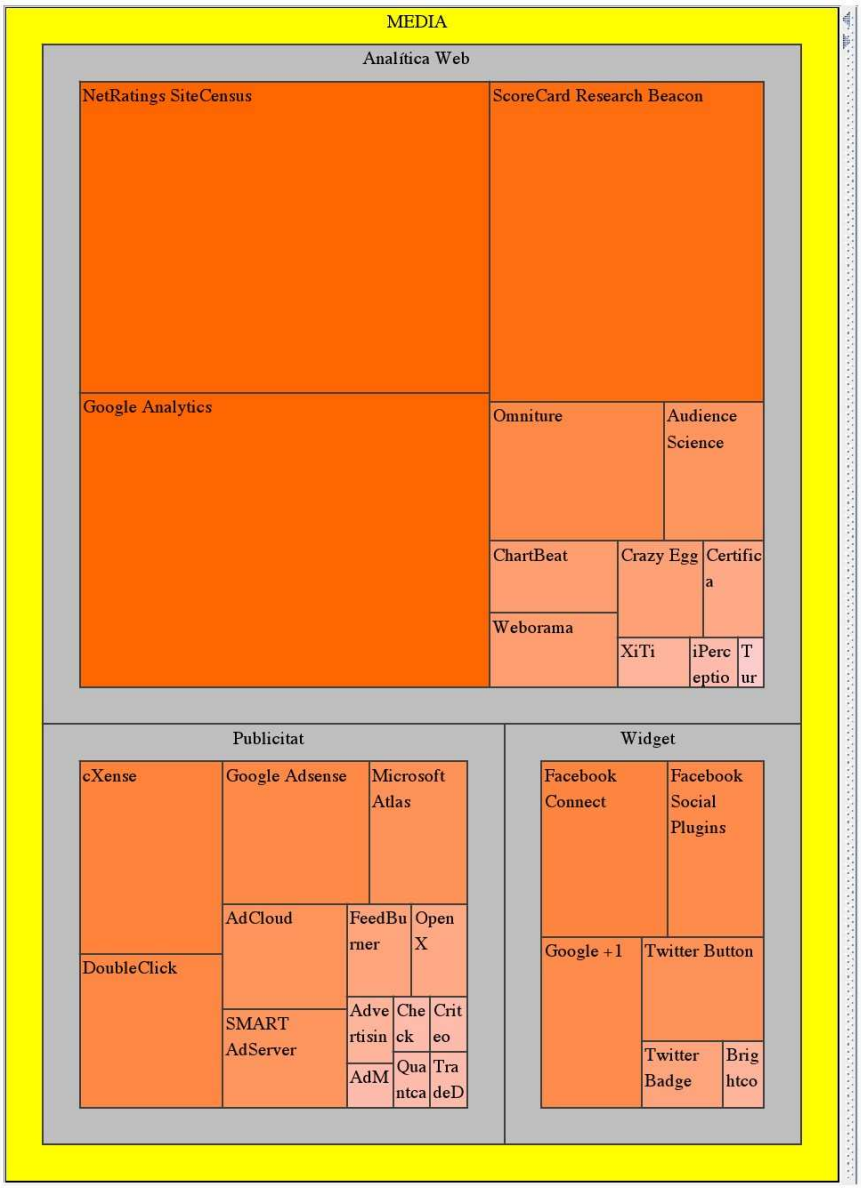
Filant una mica més prim, ens adonem que els actors inclosos dins la categoria publicitària es configuren d'una manera ben diferent en els dos espais. Si en l'espai de contractació de vols la plataforma Google Adsense es revela com l'actor dominant a molta distància dels altres, en el cas de l'entreteniment per adults això no es produeix i, de fet, s'observa un major equilibri de forces. Es pot explicar pel fet que aquesta plataforma de distribució publicitària no admet anunciants vinculats al negoci de la pornografia.

Sense moure'ns de l'àmbit del sexe *online*, convé citar un actor que només trobem en aquest espai: FriendFinder Network. Segons hem comprovat pertany a una plataforma de relacions personals *online* que abasta diferents llocs web, entre els quals un bon grapat orientats a intercanvis sexuals.

Tant a l'espai de vols com al d'oci per adults retrobem Addthis llistat com a Analítica Web.

L'espai que ofereix una configuració més singular és el dels mitjans de comunicació a la xarxa (fig. 6).

Figura 6: Treemap espai mèdia



Si bé a nivell de categories presenta una pauta similar a l'observada en els espais de vols i sexe, el grau fragmentació és molt més pronunciat. En l'apartat d'analítica web es fa palesa una situació d'equilibri entre Google Analytics i Netratings SiteCensus (propietat de l'empresa de medició i audiències Nielsen), seguits de prop de ScoreCard Research Bacon (propietat de Comscore, també

dedicada a la medició). A notar que no es tracta de presències excloents i existeix la possibilitat, i en aquest cas veurem més endavant que succeeix amb freqüència, que coexisteixin diferents eines en un mateix lloc web. Una possible explicació a la presència simultània d'aquests tres actors en els llocs web seria el fet que Netratings SiteCensus i ScoreCard Research Beacon no només s'orienten a la medició sinó que actuen com a auditors i elaboren rànquings d'audiència dels mitjans i altres informes de referència que es tenen en compte per a la planificació i contractació publicitària.

A més d'aquests tres elements, no s'ha de passar per alt el pes específic d'altres eines de medició com Omniture, Audience Science, Weborama i Chartbeat. Tot plegat ens confirma una vegada més l'espai dels mitjans com el lloc on l'activitat d'analítica web es produeix amb major intensitat.

Fixant-nos en els dispositius que integren la categoria publicitat, ens retrobem amb un nivell de fragmentació superior al de la resta d'espais i sense cap element que domini exageradament sobre els altres. Cal recordar, però, que Google pugna per l'espai a través de dos plataformes: Google Adsense i DoubleClick.

També s'observa una major quantitat d'elements de la categoria *widget* que als altres espais. A destacar el pes dels dos dispositius propietat de Facebook.

Anàlisi topològica amb grafs

En l'apartat de metodologia explicàvem que l'eina utilitzada per al procés de recollida de dades proporciona un arxiu on les dades es mostren en un format que assimila llocs web i dispositius a nodes i la presència dels elements en els llocs web a arestes (connexions). Per a representar les dades recollides en forma de graf s'utilitza el software de visualització de dades Gephi.

Les figures que venen a continuació (7, 8, 9, 10 i 11) s'han elaborat mitjançant un algoritme que organitza els elements i la seva disposició espaial forçant el distanciament dels nodes més connectats. En el nostre cas dels elements que han estat detectats en més ocasions.

La mida dels nodes, que es representen en forma de circumferències, també reflecteix la recurrència dels dispositius de manera que els cercles de major diàmetre corresponen a les eines més utilitzades.

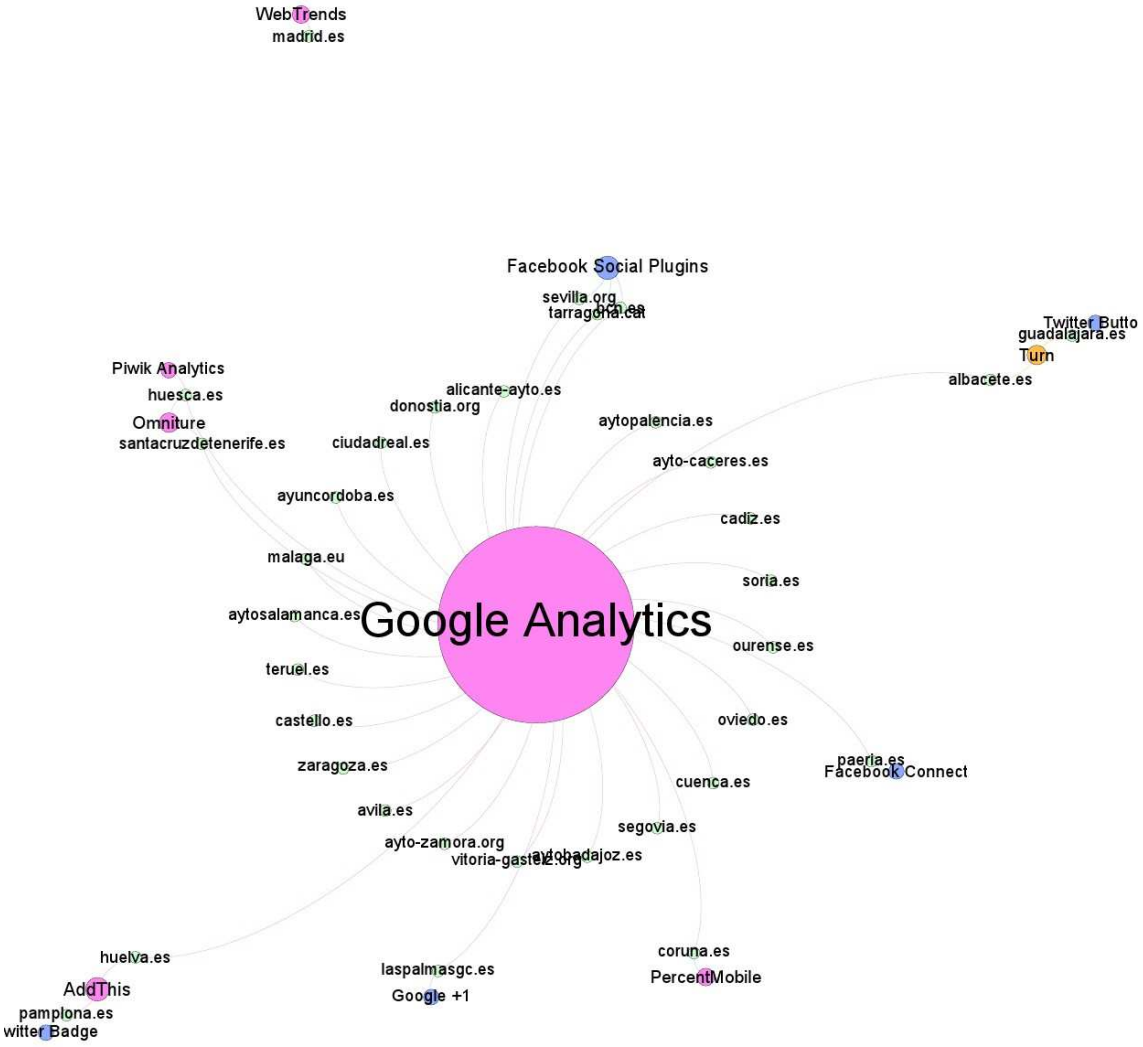
Per obtenir una visualització més llegible es força l'algoritme perquè eviti la superposició dels nodes a l'espai. Al mateix temps es modifica manualment la ubicació d'alguns elements que després d'aplicar l'algoritme apareixen molt distanciats del nucli. El motiu d'aquesta posició tan allunyada del centre del graf respon a una casuística molt determinada: elements que només es troben presents en un lloc web en el que només s'ha detectat el dispositiu en qüestió. Val a dir, que aquesta casuística només s'ha reproduït en quatre ocasions.

Les categories, o millor dit la recategorització en analítica web, publicitat i *widget*, dels dispositius es fa visible mitjançant l'ús de colors per representar els nodes. Els rosats representen la categoria d'analítica web, els carbassa, la de publicitat i el blau, els *widgets*. Per a representar els llocs web s'utilitza un verd clar. Les arestes també es mostren seguint el mateix codi de colors per tal de reflectir el destí dels fluxos de dades procedent dels llocs web.

A més del que veiem, és important tenir en compte que el graf no mostra els llocs web on la detecció d'elements ha estat nul·la.

Els grafs resultants afloren d'una forma més clara les particularitats de cada un dels espais. En una primera exploració comparativa de les figures s'observa diferents nivells de complexitat en l'estructura de xarxa.

Figura 7: graf espai capitals de província



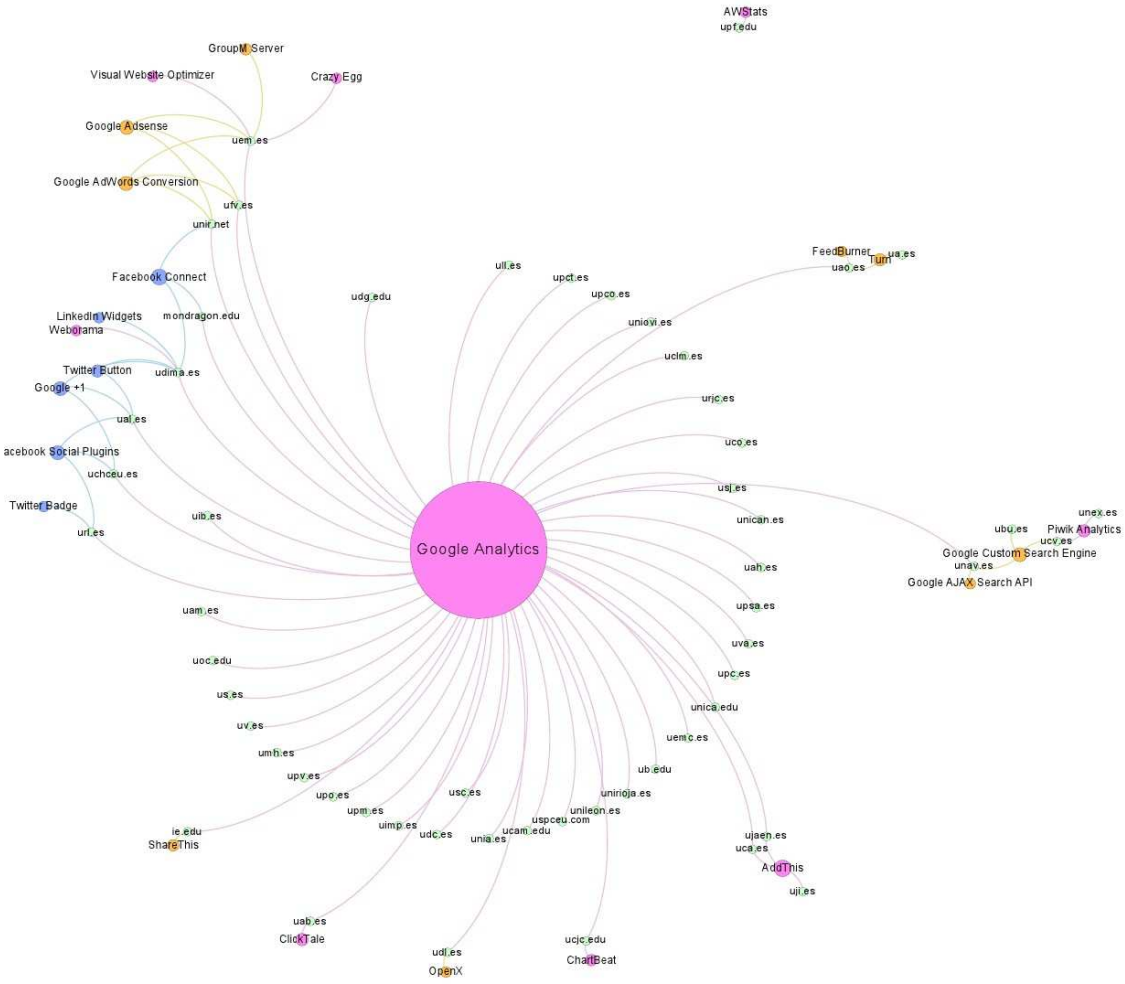
L'estructura més simple es dona a la mostra de les capitals provincials (fig. 7). Aquest fet obeeix a dos motius ja descrits anteriorment: una menor quantitat de dispositius de monitorització, i en conseqüència de fluxos de dades, i una inferior varietat en el repertori d'elements detectats.

Un altre aspecte remarcable és el paper central de Google Analytics que fa evident que és l'eina utilitzada per la immensa majoria dels portals dels ajuntaments. De fet, només apareixen tres portals que no l'utilitzen: pamplona.es, guadalajara.es i madrid.es. En aquest darrer lloc web es dona la casuística d'un únic element utilitzat que a la vegada només és utilitzat per un sol lloc web: Webtrends.

Si ens fixem en la ubicació dels nodes que representen les diferents urls de la mostra, veiem com els llocs web que només utilitzen Google Analytics es col·loquen al voltant del node principal formant una circumferència. Més allunyats del centre, apareixen aquells portals que a més de Google Analytics es serveixen d'altres instruments per recollir informació del usuaris i articular funcionalitats.

Encara més allunyats del nucli, trobaríem els llocs web que utilitzen més d'un dispositiu (que s'ubiquen encara més allunyats del centre) i/o que no utilitzen Google Analytics.

Figura 8: graf espai CRUE



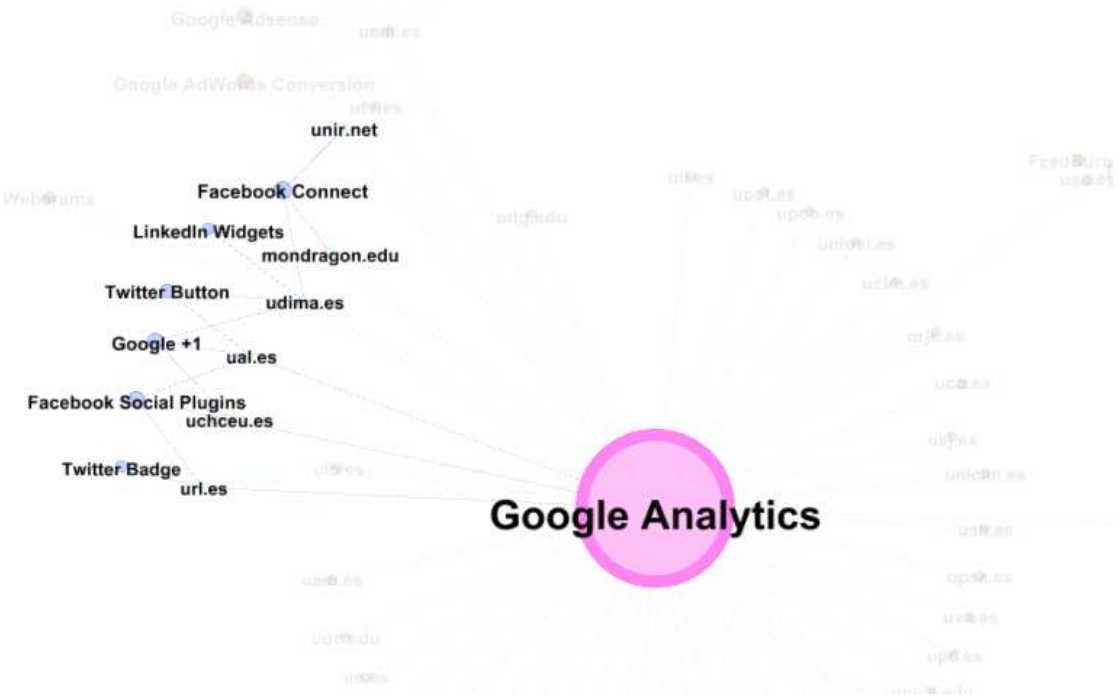
A l'espai de la CRUE (fig. 8) els elements es disposen seguint una estructura similar a la dels capitals provincials. Tanmateix, l'increment en el nombre d'elements utilitzats i les interconnexions amb els nodes dels llocs webs fan que aparegui una configuració de xarxa menys homogènia.

Google Analytics és el node més important i com a tal ocupa el lloc central del graf desplaçant els nodes dels altres dispositius a l'exterior. Els llocs web que només utilitzen aquesta eina s'agrupen al seu voltant si bé en aquesta ocasió no arriben a completar el cercle. D'aquesta disposició dels elements deduïm una diferència important en relació a l'espai d'ajuntaments: el recurs a Google Analytics com a dispositiu únic de medició es produeix amb menys freqüència.

Ells llocs web que a més de Google Analytics es serveixen d'altres mecanismes de monitorització es tornen a agrupar conformant en un segon perímetre més distanciat del node principal. A l'extrem superior esquerra del graf s'aprecia un grup de nodes verd clar (llocs web) que responen a aquestes característiques i es concentren en una superfície bastant delimitada, propera a un altre grup de nodes de color blau (*widgets*). La proximitat s'explica per l'existència d'arestes entre verds i blaus fet que propicia un tènue fenomen de clusterització. Aquest conjunt ens permet identificar les universitats que utilitzen dispositius d'interconnexió amb xarxes socials d'Internet com Facebook, Twitter, Google Plus i Linkedin.

Estaríem parlant, per tant, de les universitats que han fet una aposta més decidida per la web social. Ja sigui per servir-se de la *like economy* com a via per augmentar la visibilitat dels continguts i atraure trànsit d'usuaris o per obrir canals de comunicació amb els usuaris a través d'aquestes plataformes.

Detall cluster widgets a espai CRUE



Tot i que l'objectiu d'aquesta investigació no pretén l'anàlisi detallat de cada un dels actors, creiem necessari fer esment del cas del lloc web de la Universidad Europea de Madrid (www.uem.es) que es revela com el punt on els fluxos de dades es produeixen amb major intensitat. El repertori de dispositius inclou tres elements vinculats a l'analítica web (Google Analytics, Crazy Egg i Visual Website Optimizer) i tres més relacionats amb la publicitat (GroupM server, Google Adsense i Google Adwords Conversion). Explorant el lloc web en qüestió no trobem espais publicitaris que expliquin la presència d'aquests elements, a excepció de Google Adwords Conversion que com hem dit abans serveix per auditar amb més precisió els resultats de les accions publicitàries al cercador Google.

Detall uem.es a espai CRUE

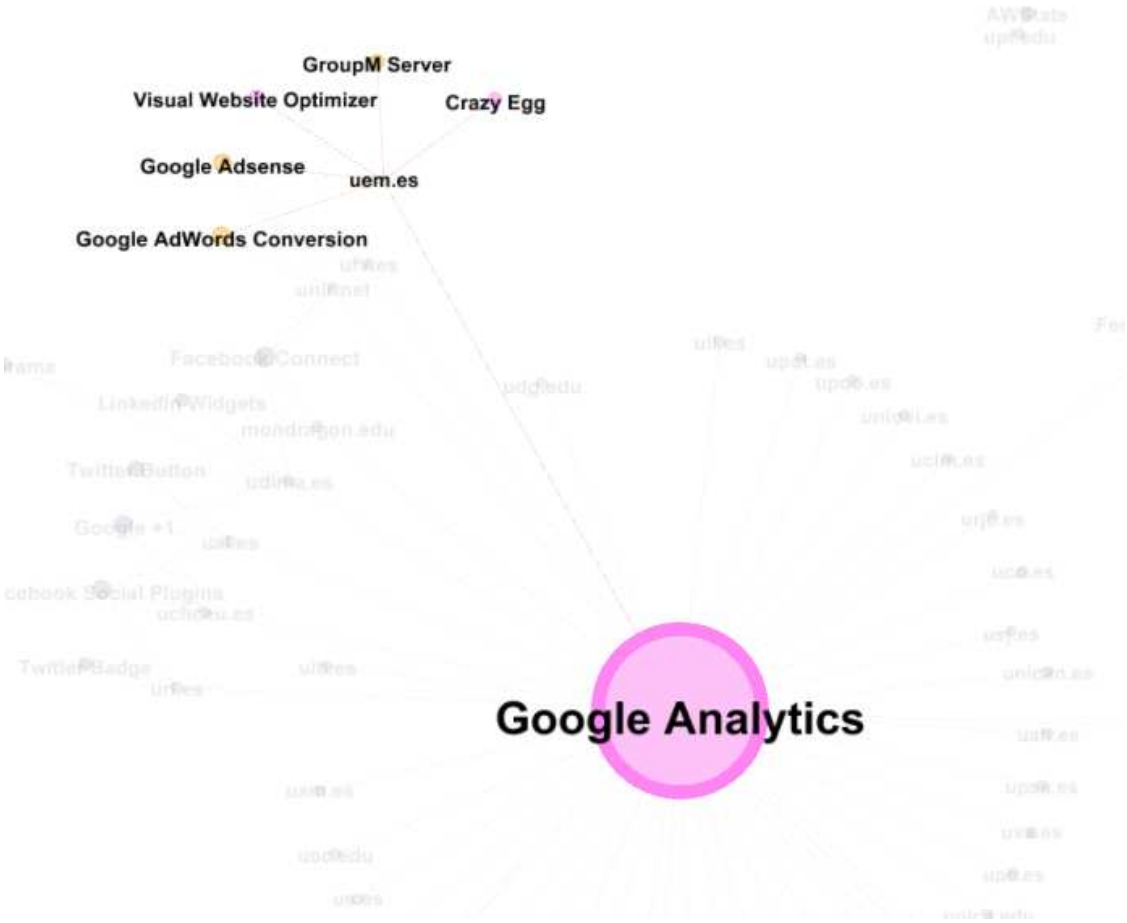
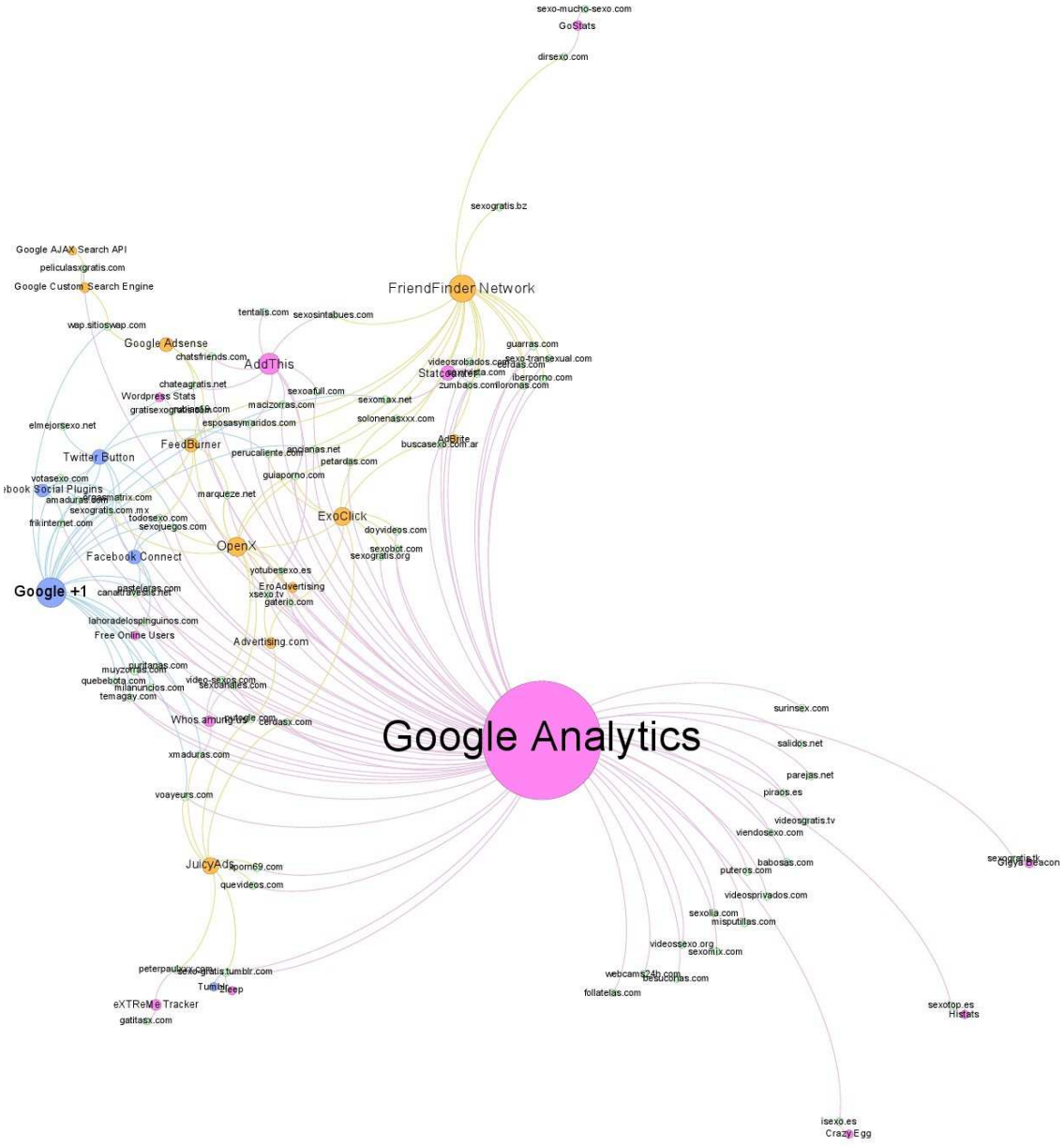


Figura 9: graf espai sexe



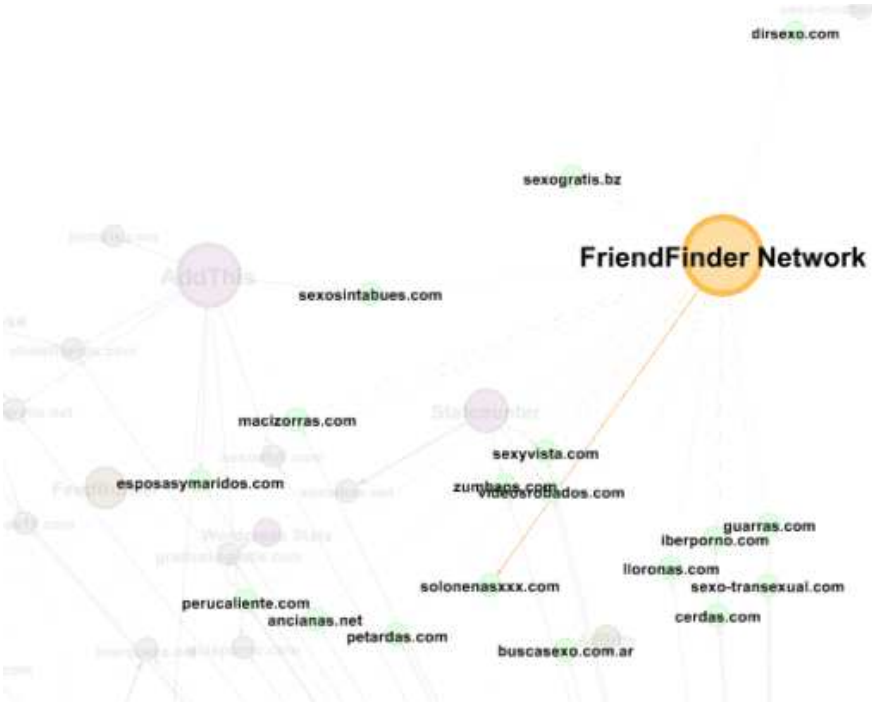
A l'espai del sexe (fig. 9) Google Analytics ocupa novament una posició destacada al nucli del graf. A la dreta, en forma de mitja lluna i orbitant al voltant del nucli, s'ubiquen un seguit de llocs web que es caracteritzen per fer servir Google Analytics com a única eina de medició.

En l'extrem oposat trobaríem la majoria de nodes que, a més del vincle amb Google Analytics, presenten ramificacions cap a diferents elements que s'ubiquen majoritàriament a l'esquerra del graf. No s'aprecia cap element que es pugui comparar en recurrència a Google Analytics però si ens fixem en la coloració dels nodes es pot comprovar una certa abundància d'elements de la categoria publicitat.

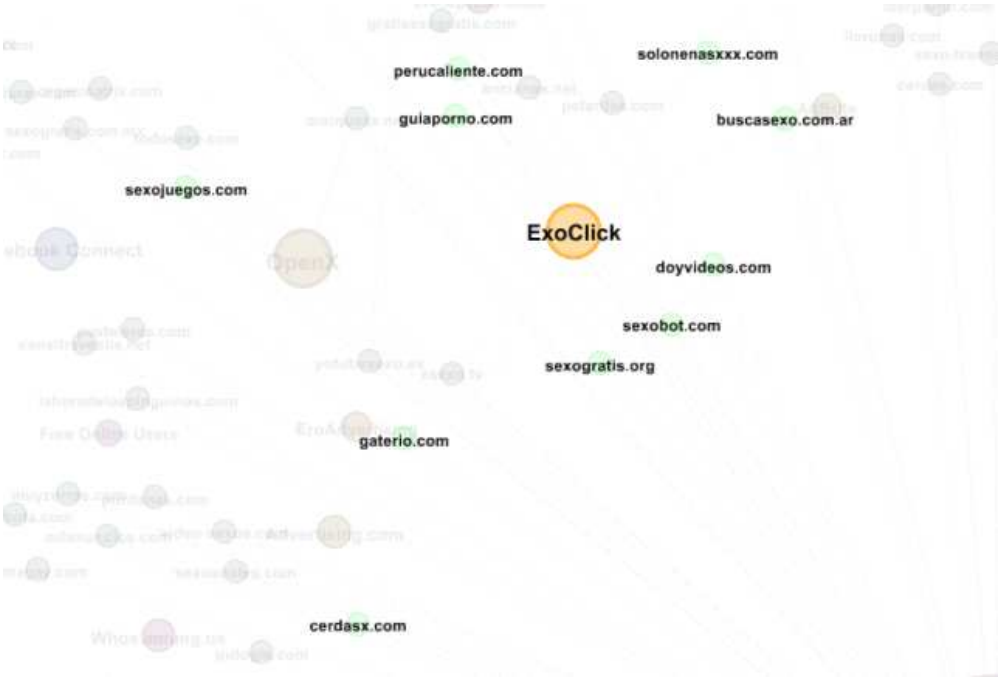
D'aquest fet es pot deduir que la difusió de formats publicitaris constitueix un model d'ingressos per molts llocs dedicats a oferir continguts d'entreteniment per adults. Tenint en compte això es pot inferir que els llocs que s'ubiquen a la mitja lluna a la dreta del node central no participen d'aquest model de negoci. Aquest fet no significa necessàriament que manquin d'afany de lucre, en aquest sector existeixen altres vies d'ingressos com la subscripció a continguts prèmium o el comerç electrònic.

La fragmentació que ja vèiem a la figura 4 ara ens permet representar l'abast de cada una de les xarxes de distribució publicitària. No hi ha cap element que gaudeixi d'una posició hegemònica però es fa evident que FriendFinder Network és l'actor més destacat d'aquest apartat.

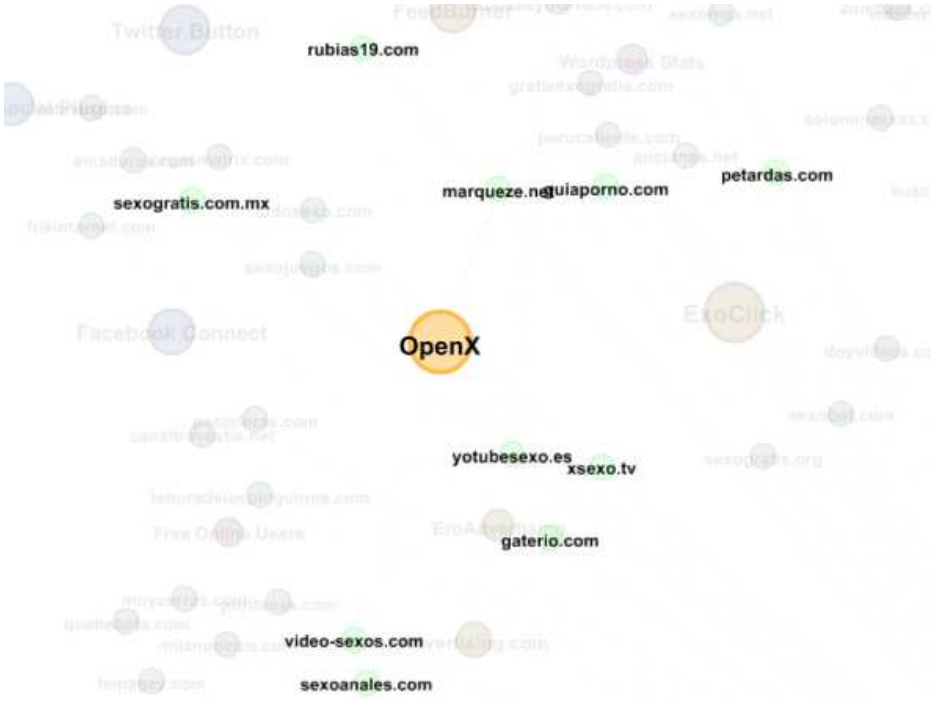
Detall presència FriendFinder Network a espai sexe



Detall presència ExoClick a espai sexe



Detall presència OpenX a espai sexe



Detall presència Google Adsense a espai sexe

Observem també una lleugera clusterització al voltant de dispositius de la categoria widget. Aquest fet, podria reflectir l'ús de les xarxes socials per a difondre continguts i atraure usuaris per part dels llocs web. De ser certa aquesta conjetura, es tractaria d'un activitat que a priori entraria en conflicte amb les polítiques de les plataformes socials pel que fa a la publicació de continguts amb sexe explícit.

Detall clúster widget a espai sexe

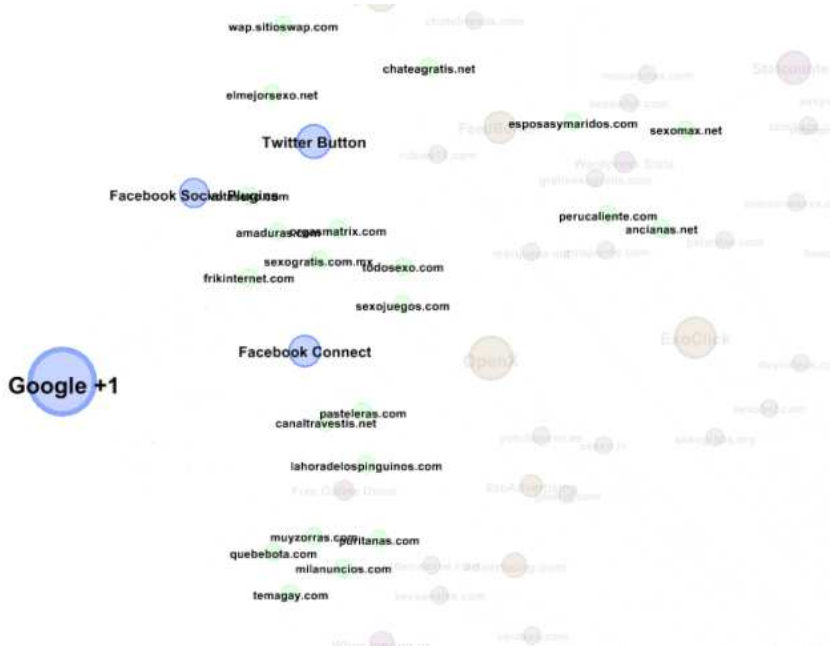
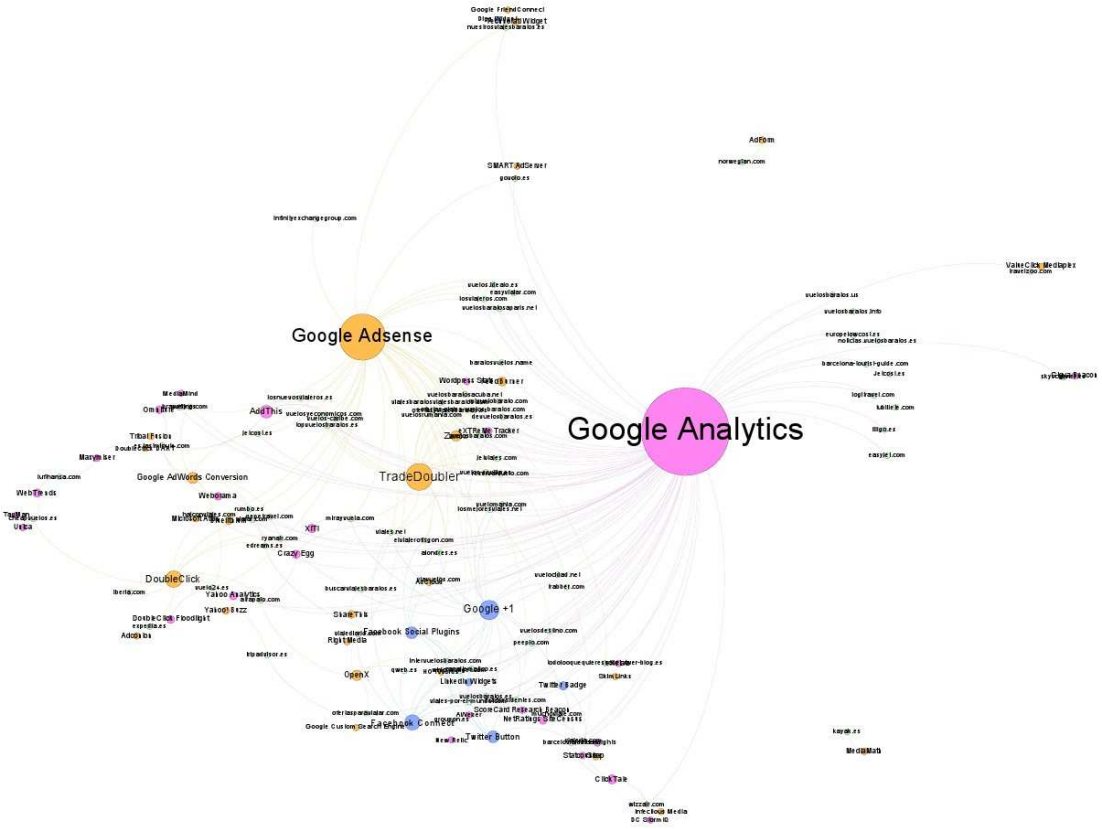


Figura 10: graf espai vols



A l’espai dels vols (fig. 10) Google Analytics és, altra vegada, el node més important. Si bé en aquesta ocasió no es situa al centre de la representació i es troba desplaçat per una gran quantitat d’elements que es situen a l’esquerra del graf conformant un mosaic de nodes de dimensions més reduïdes.

A la dreta de Google Analytics s’agrupen els llocs webs on només s’ha trobat aquesta eina. Igual que a l’espai analitzat anteriorment, es tractaria de llocs webs que no utilitzen el recurs de la publicitat per generar ingressos.

Si ens fixem en el mosaic, veurem que molts dels nodes més petits s'encabeixen dins la categoria d'analítica web. Podem llegir-ho com una tendència a fer servir altres eines d’analítica web, en molts casos de forma complementària a l'ús de Google Analytics.

Els nodes més grans de l'àrea més fragmentada corresponen a la categoria de publicitat on trobem 3 actors que destaquen sobre la resta: Google Adsense, Trade Doubler i Double Click. D'una forma molt més accentuada que en el cas del portals de continguts per adults, es posa de relleu la importància de la publicitat com a motor econòmic d'aquest espai i també el repartiment del pastís del mercat publicitari en termes de xarxes de distribució.

Detall presència Google Adsense a espai vols

Detall presència Trade Doubler a espai vols

Detall presència DoubleClick a espai vols

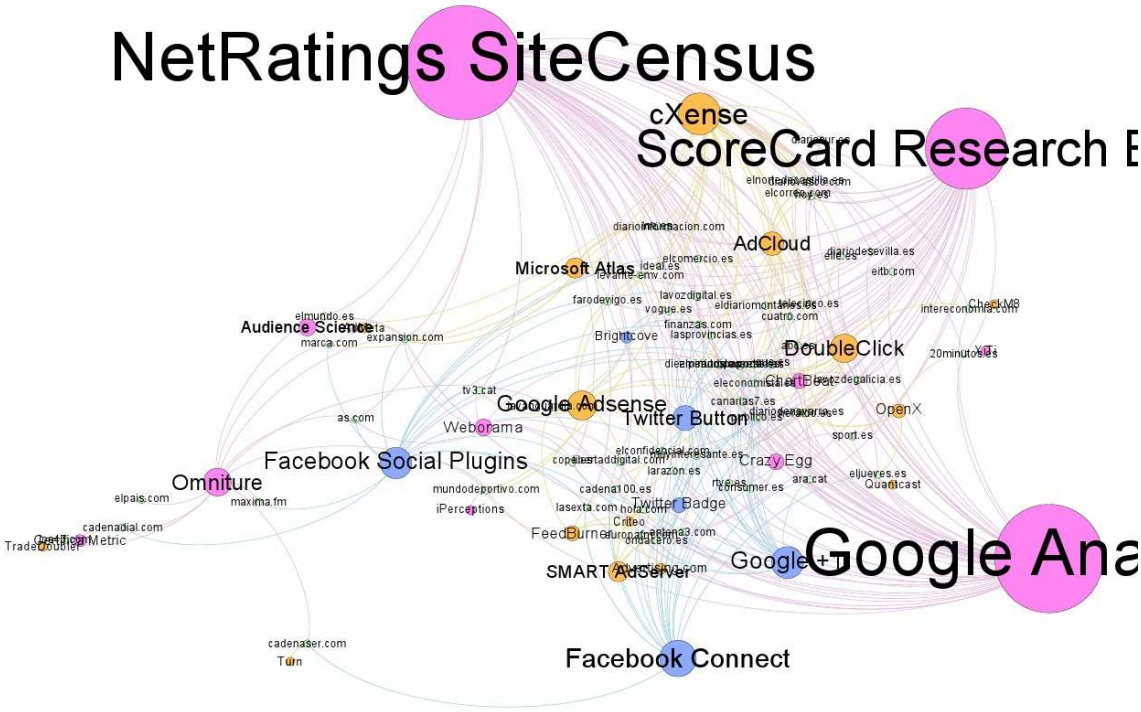
Val a dir, que Trade Doubler és una plataforma que va més enllà de la distribució d'impactes publicitaris tradicional ja que ofereix als seus clients la opció de comercialitzar productes i serveis de tercers a canvi d'una comissió.

La importància de la publicitat en aquest espai suposa un fet inesperat ja que la mostra perseguia reflectir llocs dedicats a la venda *online* de bitllets d'avió i no pas a la comercialització d'espais publicitaris. Una possible explicació rau en el mètode utilitzat per elaborar la llista de urls que, recordem, es basava en els resultats oferts per Google en relació a la cadena de cerca "vuelos baratos". Repassant la llista d'urls que integren aquest espai es fa evident que aquesta consulta no ha permès filtrar la tipologia de lloc web desitjada i trobem molts llocs web que no comercialitzen vols coexistint amb d'altres que sí ho fan. Es tractaria, per tant, d'una mostra menys homogènia i això podria incidir que l'estructura del graf sigui menys clara que en els espais vistos fins ara. En aquest sentit observem l'ús de Google Adwords Conversion com un element aglutinador d'alguns dels llocs dedicats a la venda de bitllets.

Detall presència Google AdWords Conversion a espai vols

Els elements de la categoria *widget* ocuparien el tercer lloc en recurrència i però no s'aprecia tan clarament la clusterització.

Figura 11: graf espai media

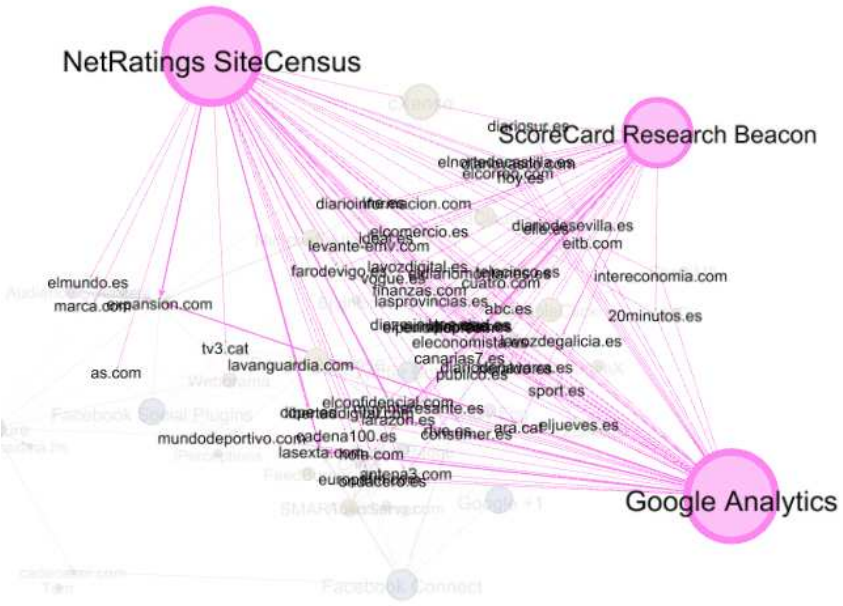


L'espai en el que conflueixen un major nombre de nodes i connexions és el dels mitjans de comunicació (fig. 11). Com a resultat d'això el graf obtingut és el que ofereix una configuració més singular i complexa.

El primer aspecte que crida l'atenció és que Google Analytics deixa de ser l'element central de l'estructura tant a nivell general com dins l'àmbit de l'analítica web. De fet, s'observa que l'element més recurrent és NetRatings SiteCensus, seguit de prop de Google Analytics i ScoreCard Research Beacon. Al tractar-se de nodes que ocupen un pes important, l'algoritme els ubica fora del centre del graf perquè la mida dels nodes actua com a força de repulsió. Al nucli, en canvi, s'ubiquen els nodes més petits i que presenten un repertori d'elements més estàndard. És a dir, que fan ús dels recursos de medició més populars.

És important remarcar que aquests tres elements comparteixen presència en la majoria de llocs web i que, per tant, no es tracta d'opcions excloents sinó de tot el contrari. Com apuntàvem abans, el fet que NetRatings SiteCensus i ScoreCard Research Beacon, a més de la recollida i l'anàlisi de dades, auditen els usuaris dels llocs web poden explicar aquest fenomen de concurrència.

Detall NetRatings SiteCensus, ScoreCard Research Beacon i Google Analytics espai media

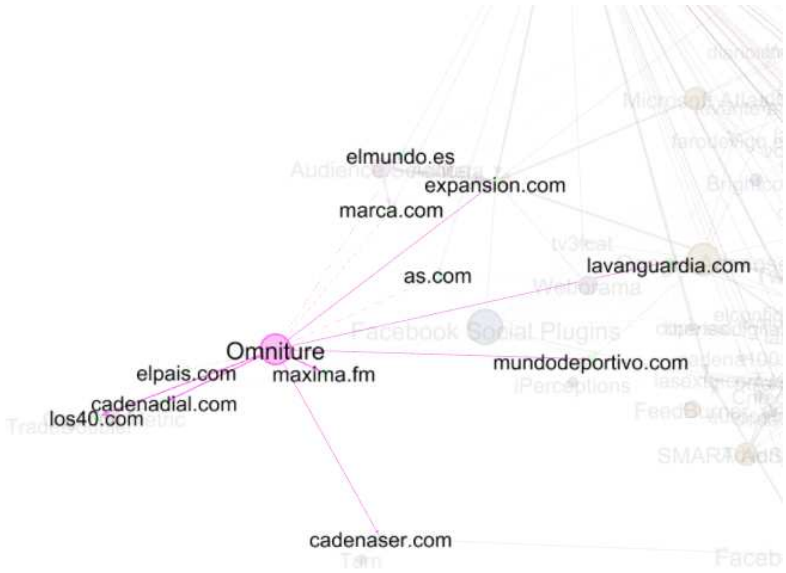


Això no exclou que trobem algunes eines d'analítica web que marcarien tendències d'ús menys estàndard. Seria el cas d'Omniture, el quart element en

importància. Aquesta eina presenta alguns trets diferencials significatius respecte a Google Analytics: és una eina de pagament i el seu ús no implica que les dades recollides es comparteixin amb ningú més.

Tot i no ser la pràctica estàndard resulta interessant observar com els mitjans digitals més importants en termes d'audiència (elmundo.es, elpais.es i marca.com) l'utilitzen i deixen de banda Google Analytics.

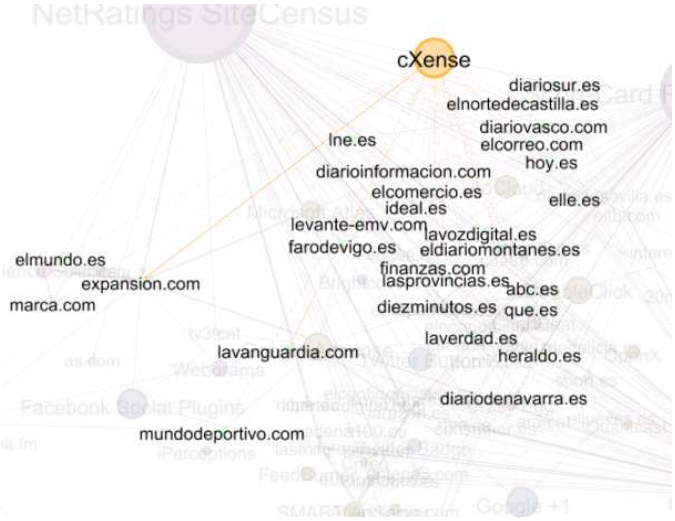
Detall Omniture espai media



En una posició més equidistant entre el nucli on es troben el gruix dels llocs webs i els grans nodes de la perifèria, trobem una profusió d'elements relacionats amb l'activitat publicitària i les xarxes socials d'Internet. Com ja anticipàvem en l'anàlisi dels *treemaps*, s'observa un presència rellevant i fragmentada dels nodes d'aquestes dues categories.

Com era d'esperar en un espai on la publicitat és el principal motor econòmic, els dispositius que articulen la distribució de les peces hi són molt presents. Tanmateix, a diferència del que vèiem en els tres grans actors de l'analítica web, la concurrència d'elements vinculats a la publicitat és anecdòtica. Aquest ens fa intuir la tendència al repartiment de l'explotació publicitària dels mitjans sota criteris d'exclusivitat.

Detall cXense espai media



Detall AdCloud espai media

Detall Google Adsense espai media

Detall DoubleClick espai media

Detall AdCloud SMART Adserver espai media

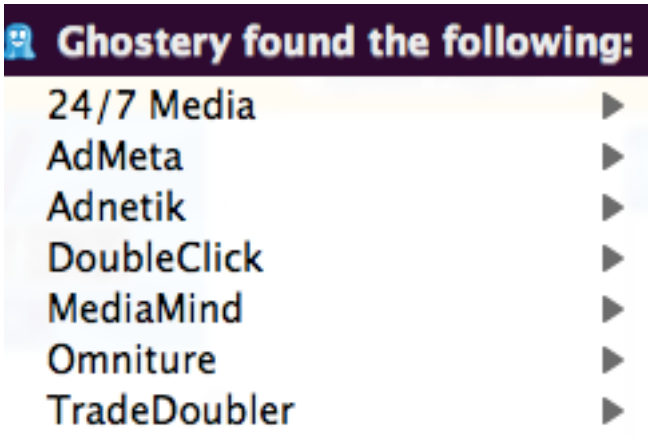
Pel que fa als elements de la categoria *widget*, es veu clarament l'aposta decidida dels mitjans de comunicació per compartir-hi els continguts. Igual que en els altres espais, cal relacionar-ho amb l'objectiu d'atraure trànsit d'usuaris i, a la vegada, obtenir una major amplificació dels continguts. En aquest sentit, sembla que l'aposta dels mitjans es decanta a favor de Facebook.

L'exploració aleatòria del graf d'aquest espai aflora un fet que ens crida especialment l'atenció: al lloc web de El País (elpais.com) només es detecta l'ús d'un dispositiu de control.

Detall elpais.com

Per descartar errors en el funcionament de Tracker Tracker es procedeix a visitar el web en qüestió amb un navegador en el qual s'ha instal·lat l'extensió de Ghostery. Sorprenentment es detecten fins a sis dispositius que han escapat a l'escrutini de Tracker Tracker.

Captura elements detectats per Ghostery a elpais.com



Per descartar la possibilitat de que la discrepància sigui deguda a la diferència entre les dates del procés de recollida de dades i les de l'anàlisi dels resultats es repeteix l'operació d'extracció utilitzant Tracker Tracker i altra volta només es detecta Omniture.

L'operació es repeteix en diferents llocs de l'espai de mitjans i tot i trobar noves diferències, aquestes són anecdòtiques i poc significatives. Tanmateix s'observa que el dispositiu utilitzat per la plataforma de distribució publicitària 24/7 Media es fa invisible al mecanisme de detecció de Tracker Tracker.

El fet més plausible per explicar tant la deficiència en la recollida de dades a elpais.es i en la identificació del 24/7 Media s'explica per la inclusió de les expressions en un arxiu independent. Que això afecti amb tant èmfasi al dispositiu 24/7 Media fa pensar que el protocol d'ús d'aquesta eina estableix el recurs a aquest mètode de programació.

11. Conclusions

Tal com s'ha dit anteriorment, un dels objectius de la recerca era cartografiar la presència d'aplicacions desenvolupades per tercers que s'utilitzen per recollir dades dels usuaris, analitzar-les i/o articular mecanismes de resposta en temps real. En l'exposició dels resultats hem vist com els mètodes i tècniques utilitzats han permès elaborar un seguit de visualitzacions del que podríem considerar un mapa dels llocs webs i els dispositius de control.

Fent servir una categorització molt bàsica i força limitada, s'han identificat 3 activitats relacionades amb l'existència d'aquests dispositius de control: la distribució publicitària, l'anàlisi de les dades de l'activitat dels usuaris i el vincle amb xarxes socials d'Internet. Tot i que al llarg de l'anàlisi hem fet èmfasi en inconsistències i deficiències d'aquesta categorització, considerem que ha estat útil per identificar, a grans trets, la natura dels elements detectats i caracteritzar els diferents espais.

El mapa resultant del procés ens mostra la presència de l'analítica web és abundant en tots els espais, fins i tot en aquells on l'activitat de monitorització és més escassa, configurant una primera capa de control.

A partir d'aquest primer nivell s'anirien articulant els mecanismes més específics que hem relacionat amb la distribució publicitària i la interoperabilitat amb xarxes socials d'Internet. Aquests delatarien una activitat de monitorització orientada a objectius més específics com la generació d'ingressos, la distribució de continguts en altres plataformes i l'atracció de trànsit d'usuaris.

Com a objectiu secundari d'aquest projecte de recerca ens havíem proposat establir una metodologia per a detectar i reflectir la presència de dispositius de control a la web.

L'elecció de les tècniques i eines utilitzades han prioritzat l'anàlisi de mostres àmplies i d'un gran volum de dades. Malgrat les limitacions que hem fet explícites al llarg d'aquesta recerca, considerem que la metodologia ha resultat raonablement útil per respondre les qüestions plantejades i assolir els objectius de la recerca.

En relació a les limitacions que s'han derivat de les eines utilitzades el detall que ens sembla més rellevant han estat els problemes amb la detecció de dispositius que s'incrusten al codi font d'elements adjacents a la pàgina HTML. En l'anàlisi del resultats hem vist que això afecta als llocs web que es decideixen per aquesta pràctica de programació però també a certs dispositius que ho estableixen com a protocol d'implementació. Preocupa més el segon dels casos perquè quan es tracta de dispositius importants en termes de penetració resulta en absències no desitjades que impedeixen fer un pas més enllà en les aspiracions de la recerca. De cara a futures investigacions fora interessant trobar una manera de superar aquest obstacle. Probablement això implicaria modificar l'eina utilitzada o buscar-ne una d'alternativa i/o complementària. Tot plegat ens ha de fer reflexionar sobre la importància –per no dir centralitat- que la metodologia de recerca juga quan es recorre a mètodes digitals d'investigació.

Un dels aspectes més reeixits de la metodologia utilitzada ha estat la possibilitat de recollir les dades de forma automatitzada i obtenir la informació tabulada. Però d'un elevat volum de dades se n'ha derivat un gran esforç en tractar i transformar les dades en formats compatibles amb el programari utilitzat per a visualitzar les dades.

Per no posar en risc la integritat de les dades, i per no complicar en excés el procés de tractament de la informació, en el procés de transformació s'han omès algunes de les variables de les que es disposava atès que no resultaven centrals per als objectius de recerca. Pensant en fases ulteriors d'aquesta recerca, s'ha de contemplar la inclusió de noves variables amb la finalitat de traçar un mapa més ric en matisos que, a la vegada, ens permetria plantejar noves qüestions i objectius més ambiciosos.

Pel que fa a les eines utilitzades s'han demostrat eficaces per a l'anàlisi de mostres àmplies. De fet, l'experiència ens ha demostrat que és a les mostres amb un major volum d'informació on aquest tipus de software rendeix a més nivell. I no ens referim únicament al fet d'obtenir visualitzacions més il·lustratives, la principal virtut d'aquest software ha estat la de facilitar l'anàlisi exploratori de les dades fent visible fenòmens molt difícils de percebre en una matriu de dades. Aquest fet ens sembla especialment rellevant de cara a

futures investigacions ja que ens obre les portes a abordar l'estudi de mostres amb més unitats d'anàlisi.

Més enllà d'aquests dos objectius generals, la recerca es plantejava respondre a un seguit d'interrogants més concrets:

La primera de les qüestions plantejades cercava identificar els dispositius de control més utilitzats. Tal com es pot observar a les diferents taules i figures, Google Analytics és l'element més recurrent a tota la mostra i també en cada un dels espais analitzats.

Aquesta posició al capdamunt del rànquing el situa a molta distància dels seus perseguidors a les 10 primeres posicions: NetRatings SiteCensus, Google Adsense, ScoreCard Research Beacon, Google + 1, Facebook Connect, DoubleClick, AddThis, Facebook Social Plugins i Cxense. En l'apartat d'anàlisi de resultats es fa èmfasi en aspectes particulars de cada un dels espais i ens semblaria una reiteració incloure'l en aquestes conclusions. En canvi, ens agradaria aportar algunes consideracions en com aquests dispositius de control serveixen per articular estructures de poder.

El rànquing no s'ha de llegir en termes de competència ja que com hem vist l'ús d'un determinat dispositiu no exclou que se'n puguin utilitzar d'altres. Però aquest grau de penetració fa evident que Google posseeix una gran quantitat d'informació relativa al consum de continguts a la xarxa. I això ens crida especialment l'atenció ja que aquesta informació li faciliten els publicadors (incrustant el dispositiu al codi font) a canvi de l'ús "gratuït" de Google Analytics. En aquest intercanvi de serveis per informació, Google no és un actor sense interessos comuns amb els publicadors. No cal anar més enllà de les 10 primeres posicions del rànquing per comprovar que la plataforma de distribució publicitària Google Adsense és la més utilitzada pels publicadors de continguts. És important remarcar que Google és molt més que un cercador de continguts a la xarxa. Els seus ingressos provenen de la comercialització d'espais publicitaris en els lloc web de la seva propietat però a la vegada també comercialitza espais publicitaris a d'altres llocs web mitjançant la plataforma Google Adsense. Gràcies a la informació que obté dels llocs web que usen Google Analytics, l'empresa de Sergey Brin i Larry Page gaudeix de la

possibilitat d'accedir a la informació d'audiències dels mitjans sense cap necessitat d'intermediació. Cal ser molt curosos a l'hora de treure conclusions precipitades respecte l'ús que Google i altres desenvolupadors de les tecnologies de control puguin estar fent d'aquesta informació. La investigació s'ha centrat en la detecció d'eines i en la seva classificació a partir de determinades funcionalitats. Esbrinar quin ús es fa de la informació recollida quedava molt lluny dels nostres objectius i de l'aproximació metodològica. Dit això, i a la vista dels resultats obtinguts, creiem que una línia interessant per investigacions futures seria l'anàlisi de les condicions contractuals que regeixen la relació entre el publicador dels continguts i el prestador de serveis de control per saber exactament amb quina moneda es paga aquest servei.

La següent qüestió sobre la que volíem indagar anava en relació al propòsit amb el que es recollien les dades dels usuaris.

La classificació dels dispositius en les categories d'analítica web, publicitat i *widgets* ens ha servit per veure com les diferents prestacions i funcionalitats de les eines es relacionen amb usos potencials vinculats a diferents processos estratègics.

Hem vist que l'analítica web és la finalitat principal de la majoria de dispositius trobats. Si ens cenyim a la definició que es fa des dels sectors professionals, aquesta activitat consistiria en la medició, recol·lecció, anàlisi i elaboració d'informes a partir de dades l'activitat dels usuaris amb l'objectiu d'entendre i optimitzar l'ús d'un lloc web. Aquesta definició té la virtut de ser concreta a l'hora de descriure el procés de tractament de les dades però també és cert que manté una ambigüitat calculada quan defineix els objectius que justifiquen dedicar temps i recursos a aquesta activitat.

La paraula entendre ens remet a l'analogia que fèiem, a l'apartat dedicat a delimitar i detallar l'objecte d'estudi, entre l'analítica web i els estudis d'audiència i recepció als mitjans. La capacitat de registrar tota l'activitat de la pràctica totalitat dels usuaris (i processar tota aquesta informació) aporta un plus important de consistència als informes que s'obtenen en l'àmbit *online*.

La paraula optimitzar està més relacionada amb la consecució d'objectius de negoci. Ser més eficaços en l'atracció de trànsit d'usuaris i aconseguir un major

consum dels continguts és fonamental per als llocs web que depenen dels ingressos publicitaris associats a la capacitat de distribució d'impactes. Al mateix temps, detectar possibilitats de millora en el rendiment de processos transaccionals també resulta d'importància cabdal per als que es dediquen al comerç electrònic.

Una categorització més granular de les eines adscrites a la categoria analítica web ens ajudaria a descriure amb major precisió totes les activitats que se'n deriven. Per exemple potser fora interessant discriminar els dispositius utilitzats per a entitats dedicades a l'estudi d'audiències (Nielsen i Comscore), l'anàlisi de les dades en temps real (ChartBeat), les que s'orienten a analitzar la usabilitat dels llocs web com (Crazy Egg, Clicktale) i moltes d'altres activitats que poden haver escapat al nostre escrutini.

Però als efectes que ens ocupen en aquesta investigació, l'ambigua definició de la Digital Analytics Association ens sembla prou raonable per definir un dels propòsits que expliquen el recurs a uns determinats dispositius.

Pel que fa a les eines englobades a la categoria publicitat, els motius semblen més evidents. Per una banda, delaten l'existència d'una activitat econòmica fàcilment identificable: comercialització d'espais publicitaris a la web i, per tant, la intenció de fer negoci. No sabem ni el volum ni la importància relativa d'aquesta activitat ja que a les diferents representacions conviuen actors molt diversos.

De l'altra, es tracta d'eines amb funcionalitats molt concretes relacionades amb la distribució de formats publicitaris: creació de perfils d'usuari, optimització del rendiment, publicació de formats provinents de xarxes externes de distribució, etc. Per aquest motiu creiem que aquí també pot resultar interessant treballar en una categorització més fina dels elements.

Continuant en la línia d'associar les categories dels dispositius amb el propòsit que activa el seu ús, toca parlar dels elements adscrits a la categoria *widget*. Tots els elements detectats que s'inclouen en aquesta categoria estan relacionats amb els mecanismes d'interoperabilitat amb xarxes socials d'Internet. Abans de cap reflexió al respecte, ens agradaria fer explícit que al llarg de tota la redacció del projecte els impulsos per canviar el nom de la

categoria han estat intensos i recurrents però hem preferit mantenir la denominació original per ser més transparents en el procés d'execució del projecte.

Tornant als motius que poden justificar l'existència dels dispositius d'aquesta categoria, cal dir que bona part d'aquests ja els hem anticipat en l'apartat dedicat a l'anàlisi de resultats. Concretament ens hem referit a l'acció de facilitar la distribució de continguts per part dels usuaris a les xarxes socials d'Internet mitjançant la publicació del contingut o un "m'agrada". D'aquesta manera els publicadors aconsegueixen amplificació del missatge fora dels seus dominis i és possible que atreguin trànsit d'usuaris procedent de les plataformes socials. A manca de saber l'eficiència d'aquesta estratègia per assolir aquests objectius, sembla que per part dels publicadors els guanys són nets si es té en compte el poc esforç que implica incrustar l'expressió de codi que activa la funcionalitat.

Potser també valdria la pena preguntar-nos en aquest cas quins són els guanys per la banda de les xarxes socials d'Internet i això ens porta a una reflexió similar a la que realitzàvem respecte a Google Analytics com a instrument per obtenir dades dels usuaris en altres llocs web. Tal com explicàvem quan delimitàvem l'objecte d'estudi, sembla que aquests dispositius també poden habilitar el flux de dades sense que es produeixi la participació activa de l'usuari. En la mesura que l'individu mantingui la sessió d'usuari a la xarxa social d'Internet activa (no se n'hagi desconnectat) aquests elements permetrien controlar-ne l'activitat en els llocs webs que incorporen aquests elements. D'aquesta manera la informació recollida permetria la construcció d'un perfil més detallat sobre les preferències i gustos de l'usuari que s'afegiria a la informació que es recull des de dins de la xarxa social d'Internet.

En aquest sentit volem destacar que els publicadors de continguts poden optar per altres expressions d'elaboració pròpia i no associades als dispositius de tercers que també faciliten la distribució de contingut i les mostres d'afecció. En aquest cas, les dades dels usuaris no estarien exposades al control per part de les xarxes socials fins que es produís la participació activa de l'usuari. D'aquesta manera el publicador preserva la informació sobre lús del seu contingut. Saber el nivell d'extensió d'aquesta pràctica seria un bon

complement per elaborar una cartografia més detallada. Per fer-ho n'hi hauria prou amb l'exploració visual de la pàgina principal de cada un dels llocs web.

La darrera qüestió plantejada es referia a l'existència d'alguna relació entre les eines utilitzades i les característiques dels llocs web.

Per tractar de donar resposta a aquesta pregunta vam dividir la mostra en diferents espais que es bastien a partir d'una sèrie de característiques. A través dels treemaps i els grafs hem pogut observar disposicions i formes ben diferents entre els espais de manera que la resposta a la qüestió seria en afirmatiu.

Hem vist que en els espais on el vincle amb interessos econòmics és menys evident, l'activitat de monitorització és menys intensa. Al mateix temps, allà on l'estratègia de continguts és més important per a la generació d'ingressos (els mèdia) el grau de control és molt elevat.

S'ha intentat presentar les visualitzacions dels diferents espais de forma gradual per fer més evidents les diferents intensitats en els fluxos de dades. Ho hem fet de menor a major pensant que d'aquesta manera facilitàvem la interpretació de les figures i no pas perquè creiem que la gradació de la complexitat tingui a veure en estadis evolutius relacionats amb les competències tècniques i professionals dels publicadors. És possible que aquest sigui un factor a tenir en compte però caldrien altres aproximacions per confirmar-ho.

Cal reconèixer que les diferències entre els espais a vegades no han estat prou clares com es dona entre els espais del sexe i dels vols. Hom podria pensar que en aquest cas l'anàlisi ens està mostrant les semblances i, de fet, és una possibilitat que no s'hauria de descartar.

Però aquests dos espais tenen en comú que la mostra s'ha confeccionat seguint un mètode idèntic (capturant les urls dels resultats de Google) que no és prou acurat a l'hora de generar mostres homogènies. Aquest és un aspecte que caldria corregir si en un futur es dona la possibilitat de repetir una investigació similar.

Tot i així, creiem que la metodologia utilitzada ha esdevingut útil per mostrar les diferències entre espais i que aquestes diferències es deuen a la orientació i finalitat dels llocs web. De ben segur que aquestes diferències es farien encara més evidents d'haver reeixit en l'intent de confeccionar mostres més homogènies.

De fet, a la vista dels resultats obtinguts estem convençuts que cal prosseguir en la via de la comparació estenent-la a més espais i contemplant la opció d'un seguiment diacrònic.

12. Bibliografia

BALACHANDER K., CRAIG E. W., Privacy Diusion on the Web: A Longitudinal Perspective. Dins de Juan Quemada, Gonzalo Leon, Yoelle S. Maarek, i Wolfgang Nejdl, editors. 18th International World Wide Web Conference (2009). ACM Press.

BEER D., Power through the algorithm? Participatory web cultures and the

technological unconscious (2009). New Media & Society 11(6): 985-1002.

BERRY, DAVID M., Philosophy of software. (2011). Palgrave MacMillan.

BOYD, D.M, ELLISON, N. B., Social Network Sites: Definition, History, and Scholarship (2007).

COLOMER, R. et al.. Societat de la informació. Noves tecnologies i Internet: diccionari terminològic. 2a ed. rev. i ampl. (2003) TERMCAT, Centre de Terminologia.

EVANS, D. S., The Online Advertising Industry: Economics, Evolution, and Privacy (2009). Journal of Economic Perspectives.

FOMITCHEV, M., How Google Analytics and Conventional Cookie Tracking Overestimate Unique Visitors dins de WWW'10: Proceedings of the 19th International Conference on World Wide Web (2010). ACM Press.

FULLER, M., Software Studies A Lexicon (2006). MIT Press.

HELMOND, A. i GERLITZ, C., The Like Economy - Social buttons and the data-intensive web (2011). Pendent de publicació.

JONES, S. G., Doing Internet Research: Critical Issues and Methods for Examining the Net (1999). Sage.

KAUSHIK, A., Web analytics: An hour a day. (2007). Sybex.

KAUSHIK, A., Web analytics 2.0. (2010). Sybex.

KNUTH, D. The art of computer programming Vol 1 (2002). Addison-Wesley.

LANG, K. et al., Efficient online ad serving in a display advertising Exchange (2011). ACM.

LESSIG, L., El código 2.0 (2009). Traficantes de Sueños.

NAKATANI, K. i CHUANG, T., A Web Analytics Tool Selection Method: an Analytical Hierarchy Process Aproach (2010). Internet Research Vol. 21, No. 2, (2011) Emerald Group Publishing Limited.

PARK, J.; KIM, J. i KOH, J., Determinants of continuous usage intention in web analytics services, Electronic Commerce Research and Applications 9 (2010). Elsevier.

PHIPPEN, A.; SHEPPARD, L. i FURNELL, S., A practical Evaluation of Web Analytics, Internet Research, Vol. 14 No. 4 (2004), Emerald Group Publishing Limited.

ROOSENDAAL, A., Facebook Tracks and Traces Everyone: Like This! (2010) Social Science Research network

ROGERS R., Operating Issue Networks On The Web (2002). Science as Culture 11(2): 191-213.

ROGERS, R., The End of the Virtual: Digital Methods (2009), Amsterdam University Press.

SANKURATRIPATI, S.; SRIVASTAVA, J. i SHANBHAG, D., Target Information generation and ad server (2006), Central Coast Patent Agency

WALL, L.; CHRISTIANSEN, T. i ORWANT, J. Programming Perl (2000), O'Reilly Media

Llocs Web

Facebook Developers Blog

http://developers.facebook.com/blog/post/108/

[consulta: juny 2012]

Electronic Frontier Foundation:

http://w2.eff.org/Censorship/Internet_censorship_bills/barlow_0296.declaration

[consulta: juny 2012]

Ghostery

http://www.ghostery.com/about

[consulta: juny 2012]

Web Analytics Association:

http://www.digitalanalyticsassociation.org/?page=aboutus

[consulta: juny 2012]

Wikipedia contributors, "Client–server model," Wikipedia, The Free Encyclopedia:

http://en.wikipedia.org/w/index.php?title=Client%E2%80%93server_model&oldid=492876069

[consulta: juny 2012]